Journal of Phonetics 55 (2016) 38–52

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Research Article

P-centres in natural disyllabic Czech words in a large-scale speech-metronome synchronization experiment

Pavel Šturm*, Jan Volín

Institute of Phonetics, Charles University in Prague, n. J. Palacha 2, Prague 1, 116 38, Czech Republic

ARTICLE INFO

Article history:
Received 12 March 2015
Received in revised form
17 November 2015
Accepted 20 November 2015
Available online 17 December 2015

Keywords:
Czech
P-centre
Speech rhythm
Syllable
Synchronization

ABSTRACT

The difficulty of pinpointing a specific event within words that would correspond to the p-centre is well known. The current experiment, investigating the position of p-centres in Czech, aims to replicate the findings from English and several other languages, and substantially increase the range of phonotactic types and the number of participants. In a speech-metronome synchronization task, 24 subjects pronounced a set of 37 natural disyllabic Czech words of differing complexity at two metronome rates. The beginning of the first vowel (V1) and the moment of the fastest increase in energy within the first syllable were the most consistent synchronization points, but the p-centre occurred earlier than at the V1 initial boundary. Synchronization intervals were significantly influenced by the complexity of the syllabic onset: the p-centre was positioned earlier (further from the V1) as more consonants were included in the onset. The effects of vowel length and final coda were also present, but weaker. In addition, various aspects of human musicality were found to correlate with the ability of speakers to synchronize their articulations with an isochronous auditory sequence.

1. Introduction

1.1. Rhythm in speech

Despite the fact that rhythm is a ubiquitous and familiar phenomenon in both the animate and inanimate world, the role of rhythm in speech has only recently been duly acknowledged. Similarly to other types of human behaviour from breathing and walking to music, dance and collective chanting, speech exploits rhythmicity in order to be more effective, albeit less explicitly. It is common that speakers adjust their temporal organization of speech so that it more closely resembles that of their conversation partner (Cummins, 2009), which may facilitate the communication process and indicate willingness to cooperate. Moreover, it has repeatedly been demonstrated that a natural rhythmic flow of speech is easier to process by the brain, leading to shorter reaction times, than uncommon or unpredictable patterns of prominence contrasts (e.g. Buxton, 1983; Quené & Port, 2005). A theoretical explanation is offered by the neural resonance model of Grossberg (2003) which states that the formation of a percept of any object (syllable, word, etc.) requires highly synchronized activations of neural assemblies in the brain that pertain to the input neural representation on the one hand (analysis of the incoming signal), and the expectational representations on the other (derived from experience and analysis of the context). The argument is discussed more specifically with regard to rhythm in Ghitza and Greenberg (2009). It is no longer surprising, then, that good public speakers tend to be – consciously or subconsciously – more rhythmical than less skilful orators, as such a performance demands less mental effort on the part of the listeners (Kohler, 2009). In a similar vein, specific kinds of rhythm may also contribute to speakers' attractiveness or to the credibility of their propositions (Knight & Cross, 2012).

To put speech rhythm into a larger perspective, numerous experiments on timing control from across disciplines have shown that if we are dealing with recursive actions, rhythmic movements are executed more easily than arrhythmic ones (Cummins, 2009; Kelso, 1995; Port, 2003; Repp, 2005; Turvey, 1990). Regular is not only easier, but it also allows for coordination of actions. For instance, Kelso (1984) asked subjects to oscillate their index fingers to the left and right at various speeds as determined by metronome pulses. The task proved to be most effortless when both fingers moved towards or away from each other, irrespective of tempo,

* Corresponding author. Tel.: +420 221 619 250.
E-mail address: pavel.sturm@ff.cuni.cz (P. Šturm).

0095-4470/$ - see front matter © 2015 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2015.11.003

whereas other phase relationships were more unstable, with the exception of the synchronous movement of both fingers either to the left or right (but only in the slow tempo). These two modes of wagging seem to work as attractor states towards which the subject's performance converges. Although to a large extent independent, the two hands prefer to work in coordination. More importantly, comparable results were obtained when the subject performed a cyclic action while watching someone else, so it is not merely a result of the physical link between the two hands (Port, Tajima, & Cummins, 1996). Further, applying such principles to speech, Cummins and Port (1998) discovered that in a speech cycling task analogous discrete coordinative patterns emerge. The internal timing of a phrase that was repeated in time with a metronome was not completely autonomous; rather, in that experiment certain stressed syllables inclined to 1/3, 1/2 and 2/3 of the phrase repetition cycle, revealing a harmonic relationship.

Researchers have approached rhythm in speech from various angles. The so-called rhythm metrics, capable of capturing some interesting (structural) differences between languages, individuals or speaking situations, have been used extensively in recent years, although it has been rightly recognized that they provide little information about rhythm as such (e.g. Kohler, 2009). Crucially, these metrics fail to capture the fact that rhythm operates beyond pure duration and, moreover, that rhythm is a perceptual phenomenon (Cummins, 2009; Kohler, 2009; Lee & Todd, 2004; Lehiste, 1977; Volín, 2010). It was acknowledged early on that an acoustically regular arrangement of stressed syllables does not result in perceived isochrony (Lehiste, 1977). If we subscribe to the perceptual view of rhythm, this finding must be neither surprising nor discomforting. On a more general level, the lack of monotonous chains of rhythmic units is to be expected in conversational speech because our speech behaviour is ultimately guided by the requirements of the content (e.g., the structuring of the idea to be expressed; Volín, 2010). More specifically, the metrical "beat" the listener is assumed to perceive in individual words is simply not associated with the acoustic onset of the word or syllable, but its position is determined by a number of acoustic and psychoacoustic factors. The moment of the perceptual emergence of the syllable in the mind of the listener was termed the "P-centre" (Morton, Marcus, & Frankish, 1976).

1.2. Investigation of p-centres

Early inquiries into the nature of p-centres investigated the position of rhythmic stress beats in English. Although the subjects in Rapp's (1971) study repeated nonsense words to a regularly occurring pulse while Allen's (1972) experiment included finger tapping to the presumed beat of a specified syllable in a sentence, in both paradigms the point of synchronization was located near the acoustic onset of the vowel. More importantly, it shifted backward in time (i.e., further ahead from the vowel) in direct proportion to the prevocalic consonantal duration. This finding was corroborated by the classic study of Morton et al. (1976) who found that sequences of digits that were aligned perfectly with respect to their acoustic onsets were not perceived as rhythmical. When instructed to adjust the intervals between the digits for better rhythmicity, the listeners introduced considerable variation into the inter-onset intervals. Taken together, these experiments revealed that departures from acoustic isochrony are systematic and vary in tandem with the duration of the prevocalic material, but nevertheless failed to pinpoint any acoustic marker corresponding to the p-centre, such as the acoustic onset of the word or of the stressed vowel.

Several factors have been identified to play a major role in the determination of p-centres. The duration of the (acoustically salient) prevocalic consonantal portion of a syllable is positively correlated with a backward shift of the p-centre (Allen, 1972; Fowler, 1979; Fowler & Tassinary, 1981; Marcus, 1981; Rapp, 1971). Cooper, Whalen, and Fowler (1986) confirmed that the duration of the onset is the key factor, not its phonetic identity; however, the results of Harsin (1997) based on a wider range of segments suggest that phonetic categories do have an effect on p-centre placement via their acoustic properties, as the size of the effect of onset duration differed for different initial consonants. The relevance of the stressed vowel in terms of its duration was reported by Marcus (1981) and by Fox and Lehiste (1987a), who concurred that longer vowels are associated with later p-centres. The latter investigated tense and lax vowels in English monosyllables and found that it was specifically duration and not vowel quality that contributed to the effect. Marcus (1981) and Fox and Lehiste (1985) reported a significant effect of the final consonant, as did Cooper, Whalen, and Fowler (1988) for both the nucleus and the coda. Finally, Fox and Lehiste (1987b) examined how the location of the stress beat is affected by adding an unstressed prefix or suffix to a monosyllabic word. In both production and perception, an unstressed suffix pulled the p-centre later into the stressed vowel, whereas a prefix exercised a much larger shift in the opposite direction. This demonstrates above all the necessity of investigating more complex stimuli in addition to the simple sequences of stressed monosyllables that have hitherto been the main focus of p-centre research.

A perceptual aspect of the p-centre phenomenon is advocated by Pompino-Marschall (1989), who developed a psychoacoustical, rather than an acoustical, model of p-centres that is based on loudness functions within critical auditory bands. He agrees with Marcus (1981) in assigning greater impact to onsets rather than to the rest of the syllable, but differs from him in establishing interactions between onsets and vowels, and vowels and codas. Moreover, he found that the amplitude envelope of the syllable plays an important role (similarly to Howell, 1984, 1988a,b). In a similar vein the study of Harsin (1997) points to the role of low-frequency energy modulations in the acoustic signal, in particular the velocity peaks at the C–V transitions. P-centres thus seem to be affected not only by the duration of its constituent segments but also by their energy distributions and, by inference, their phonetic identity, which contrasts with the already mentioned data of Cooper et al. (1986) or Fox and Lehiste (1987a). In any case, it is evident that properties of the whole stimulus participate in determining the location of p-centres.

Articulatory correlates of p-centres were investigated in the work of Carol Fowler and her colleagues (Fowler, 1979, 1983; Fowler & Tassinary, 1981; Tuller & Fowler, 1980). Although restricted to only a few subjects and to simple monosyllabic items (e.g., sequences of /sæd mæd stæd stræd/ etc.), her results support the previous findings in a variety of production and perceptual tasks. It is clear that in speech production speakers synchronize some event within a syllable, and presumably within its consonantal onset.

Fowler (1979) suggested that the p-centre might be associated with the articulatory activity related to the vowel gesture, such as the moment of its initiation or of reaching the target vocal tract shape. This would be reflected in the signal as coarticulatory influence of the vowel on the preceding consonant. Listeners are then hypothesized to exploit acoustic information about the articulatory timing of vowels when judging rhythmicity.

However, as pointed out by Marcus (1981), this strictly articulatory position is difficult to maintain. His perceptual experiments, in which listeners adjusted the position of the final stimulus so that it would be perceived as isochronous with the three preceding stimuli, revealed not only effects of the onset consonant characteristics on the location of p-centres, but also a significant influence of the vowel itself and even of the coda consonant: a longer vowel and/or a longer final consonant led to a p-centre shift towards the centre of the vowel. Marcus's acoustic model thus incorporates two competing forces (syllable onset vs. rime) exercising influence in opposite directions. Similarly ambiguous results about articulatory correlates of p-centres were presented by de Jong (1994). In his experiment acoustic parameters correlated with perceptual p-centre adjustments equally well as X-ray microbeam data, with considerable individual variation. Patel, Löfqvist, and Naito (1999) did not find a consistent articulatory cue, either. It thus seems untenable to claim that p-centres coincide with some aspect of the vocalic gesture given that the whole stimulus affects the p-centre perception. It is argued that p-centres necessarily reflect both articulatory and perceptual factors, emphasizing the view of "speech as a medium of communication between speaker and listener" (Marcus, 1981: 255). Needless to say, human perception and production are massively interconnected, and their separation is mostly a purely descriptive device. From this point of view, it is convenient that the term "p-centre", originally denoting perceptual centres, can also stand for produced centres of the syllable.

Another concern relates to the universality of p-centre position and speech timing in general. As pointed out by Barbosa, Arantes, Meireles, and Vieira (2005), attention has been restricted mainly to the Germanic languages; indeed, most of the experiments reported above were performed on English or even on non-speech stimuli. Cross-language comparison is therefore highly desirable, especially in light of the fact that the rhythm type of an individual language predestines its users to specific strategies in speech perception, speech acquisition and, naturally, speech production (see, for example, a review in Cutler and Otake, 2002). However, Hoequist (1983) confirmed that the speakers of Spanish and Japanese behaved as those of English: all three language groups demonstrated comparable p-centre effects in a production task. Although Fox (1987) likewise showed that English and Japanese listeners were similarly affected by the duration of the vowel, which suggests universality, the two groups differed, albeit slightly, with respect to manipulations of the coda consonant. This might reflect an influence of language-specific properties pertaining to syllable structure. Recently, Barbosa et al. (2005) reported valuable p-centre data on Portuguese, Volín, Churaňová, and Šturm (2014) on Czech, and Chow, Belyk, Tran, and Brown (2015) on Cantonese. There is a clear demand for extending the description of p-centre behaviour to a broad variety of languages.

1.3. Research questions

The research presented above gives rise to several objectives in our study. First, there is an urgent need for experiments that employ a sufficient number of subjects. Barbosa et al.'s (2005) results were based on one subject only, which immediately arouses suspicions about the validity of the conclusions for the population of Brazilian Portuguese speakers as such. Marcus (1981), Fox and Lehiste (1987a), Cooper et al. (1986) and Pompino-Marschall (1989), to name just a few, used three or four subjects. The only representative study in this respect is Chow et al. (2015) with 23 participating speakers, who, as the authors admit, "behaved quite differently from one another" (Chow et al., 2015: 63). It is unfortunate that these differences were not analysed and reported in the article because, as Pompino-Marschall (1989) notes, the rhythmic behaviour of subjects can be extremely diverse. Therefore, one of the aims of our experiment was to evaluate the rhythm aptitude of our subjects and test the degree to which it correlates with their musical background (see, e.g., Repp, 2010 for an investigation of sensorimotor synchronization skills by musically trained and untrained subjects). This naturally entails an adequate size of the speaker sample.

A second recurrent objection concerns the nature of the stimuli used in experiments. The majority of p-centre research was based on simple monosyllabic words or – like Fowler and Tassinary (1981) – even on nonsense syllables. Although this is fine at initial stages of research, 30 years later it is necessary to examine a more complex material. Obviously, it is not possible to use conversational speech in a speech-metronome synchronization experiment, but subjects can at least be confronted with real, commonly used words to increase the ecological validity of the task. The obvious penalty for selecting real words is a lesser degree of control over the structural properties of the items.

Furthermore, the research should be extended to polysyllables since rhythm is by definition a matter of alternation of more and less prominent syllables. Czech has relatively few monosyllabic words (Bartoň, Cvrček, Čermák, Jelínek, & Petkevič, 2009) which, in addition, tend to combine with other words to form polysyllabic stress groups in connected speech (Churaňová, 2013). However, the most important argument is that, as several experiments reported above indicated (e.g., Fox & Lehiste, 1987b), p-centre location seems to be affected by properties of the whole stimulus. We thus employed meaningful words of two syllables with the general aim of testing natural words, while one of the specific aims was to investigate the effect of the structural properties of the second, unstressed syllable on the position of the p-centre.

Finally, as was already mentioned, the current experiment can also be viewed as a response to the frequent complaint that apart from some Germanic languages (predominantly English) we lack informative data in the domain of p-centres (e.g., Barbosa et al., 2005; more recently Chow et al., 2015). The latter study is especially useful as it deals with Cantonese, a tonal language with a particularly restricted syllabic structure. Czech, on the other hand, has a relatively complex syllabic structure, but consonantal clusters are formed according to different principles than in English or German. In addition, Czech also differs from the Germanic

languages in the fact that the presence of lexical stress is usually not associated with vowel lengthening. Preliminary results on the Czech language were already reported in Volín et al. (2014). The present study continues to map the behaviour of Czech speakers and provides a more complete account of the data, especially as regards the range of phonotactic types investigated.

2. Material and methods

2.1. Stimuli

The experiment was aimed to test the behaviour of a wide array of phonotactic structures. For reasons outlined above, we used existing disyllabic Czech words rather than nonsense words or existing monosyllables. The targets contained the vowel /ɛ/ or /ɛː/ in the first (stressed) syllable and the vowel /a/ or /aː/ in the second (unstressed) syllable, resulting in four combinations of vowel length. Other Czech vowels were not included due to their relative infrequency, occurrence only in foreign words or a salient qualitative difference between the short and long counterparts. Moreover, the inherent difference in individual vowel durations could complicate the design. The stress was always on the first syllable, which is a canonical feature of Czech polysyllabic words. The onset of the first syllable consisted of one or two consonants in the majority of cases (C × CC), but for the /ɛː aː/ vowel pairing we also added a few words with a three-consonant onset (CCC). The segmental content comprised three major classes (plosives, fricatives, liquids). The intervocalic position always consisted of only one consonant (an obstruent or a nasal). Finally, the second syllable was either open (no coda consonant), or contained a single coda consonant (an obstruent or a sonorant). Such a design yielded 18 phonotactic types (8 for C onsets, 8 for CC onsets, 2 for CCC onsets), with 37 target words in total (see the Appendix).

2.2. Subjects and musical background

The synchronization task was completed by 24 native speakers of Czech (18 females and 6 males, aged from 20 to 32, with mean age of 23.9). The participants were mostly students at Charles University in Prague, and reported no speech or hearing impediments. They received financial compensation for their involvement.

After the experiment each participant was asked to fill in a form that clarified his or her musical background. The questionnaire comprised ten questions about musical training, current and past musical activities (including specification of frequency of practicing), motivation, listening habits, and attitude to dancing. The qualitative and quantitative data were converted into numerical scores that were later added up, under various assigned weightings, into a single *musicality score* for each participant (see Section 3.4).

2.3. Task procedure

The subjects were seated in a professionally sound-treated recording booth during the experiment. In each trial they heard a series of twelve metronome pulses in the headphones while a target word appeared on the computer screen. Their task was to pronounce the given word in synchrony with the beats until the metronome stopped and was replaced with a short stretch of soft music. Articulation began with the fifth pulse since the initial four pulses served as a lead-in for steadying the speakers' attention. The subjects were specifically instructed to synchronize their repeated articulations with the isochronous sequence of pulses so that each pulse was aligned with the first syllable of the word spoken in isolation. A clear and "natural" pronunciation was demanded, as natural as it was possible in the experimental conditions (the subjects were explicitly asked not to chant or recite the words). In order to protect the data from initial and final lengthening (Klatt, 1976), as well as from effects of accommodation at the stimulus beginning and anticipation of the stimulus end, the first and last realizations were omitted from analyses, leaving 6 repetitions for each item (henceforth *sextets*).

Two tempi were employed because it is hypothesized that listeners use different processing modes depending on pace (Kohno, 1995; compare also different results for the various tempi used in Barbosa et al., 2005). The metronome pulses thus appeared at the rate of 70 bpm in the "slow" version (857 ms intervals) and at the rate of 90 bpm (667 ms intervals) in the "fast" version. The two values were chosen to induce production rates of about 2-3 and 3-4 syllables per second since the participants always produced one disyllable word per pulse. (It should be noted that Chow et al. (2015) employed even lower values, namely the rate of 60 bpm.) Since mechanical metronomes and even computer metronomes can vary in precision (Quené, 2007: 354) we decided to record a sample metronome sound and copy it at the required, entirely identical intervals for the duration of the trial. The subjects thus listened to audio files instead of a real metronome.

In addition to the 37 target words which appeared in both tempi, filler items (mono- or trisyllabic words with no restriction on the segmental content) were used to provide variation to the task and prevent subjects from lapsing into monotonous behaviour. Trials were presented in DMDX (Forster & Forster, 2003) in six blocks with a one-minute rest in between, during which the subjects chatted with the experimenter and could stretch their bodies or refresh themselves with a drink. Tempo was used as a blocking factor. Moreover, the experiment was self-paced, the next trial being selected by pressing a button. All trials were randomized within the blocks. The duration of the session was approximately 40 min.

2.4. Segmentation and extraction of data

Two channels were used simultaneously to record the session, one for the metronome pulses and one for the spoken production. A notebook with a high-quality external soundcard (EMU USB0202) and ASIO drivers played back audio files containing the metronome pulses; the main output was being recorded directly into a computer (in the left channel), while the subject listened to a direct monitor output (with zero temporal latency) over the headphones. The subject's speech was being recorded into the right channel using a condenser microphone. In order to establish to what degree both channels were synchronized we replaced headphones with loudspeakers for a test run and recorded the metronome pulses directly (L channel) and over the microphone (R channel). From a visual inspection of the waveform there seemed to be at most 1-ms difference between the two channels.

The stereo recordings were processed and annotated in Praat (Boersma & Weenink, 2013). An in-house script based on the detection of an intensity threshold in the metronome channel was used to determine the position of the recorded metronome pulses, which allowed us to identify the location of the metrical beat (p-centre) for each produced word (but see Discussion with regard to p-centres and metronome pulses). Individual trials were then segmented into words (six tokens in each trial) and phones using the Prague Labeller algorithm based on HMM and forced alignment (Pollák, Volín, & Skarnitzl, 2007; see Volín, Skarnitzl, & Pollák, 2005 for a comparison with human evaluation); these pre-processed segment boundaries were manually checked and corrected where necessary. Some of the segmentation rules that we followed during the latter phase included:

- *vowel-obstruent boundaries* (or vice versa): the boundary was placed at the point where full formant structure began/ceased to appear; i.e., periodic glottal pulses were not a sufficient reason to assign this portion to a vowel; thus, our segmentation of vowels is rather conservative;
- *second vowels* were sometimes accompanied by breathy phonation which gradually changed into the final voiceless consonant ([t ʃ r̥]); in these cases, the boundary was placed after some clear turning point in the spectrum, or – when no solution was motivated by the signal – at midpoint of the transitory region;
- *obstruents* are characterized by a lack of formant structure and by reduction of energy especially in the lower frequency regions; fricatives have in addition friction noise;
- *nasals* are associated with reduced energy in higher frequencies, and have a strong nasal murmur in lower frequencies (a nasal formant); there is an abrupt change in amplitude and waveform shape at boundaries with vowels; NV transitions often include explosion;
- *laterals* have full formant structure but formant frequencies are lower compared to vowels; also, laterals are associated – like nasals – with reduced energy in higher frequencies; the LV transition is often marked with abrupt increase in amplitude.

It should be noted that contrary to most p-centre investigations, initial voiceless stops were segmented as full consonants, and not truncated to their release burst phase. Consequently, the boundary placement in these sounds with initial silence was arbitrary (but consistent throughout).

After discarding seven trials (42 word tokens) with incomplete or disrupted items (i.e., missed repetitions, dysfluencies), the final data set was extracted from 10,608 word tokens (rather than the predicted 10,656: 24 speakers $\times$ 2 tempi $\times$ 37 target words $\times$ 6 repetitions). Using Matlab (The MathWorks, Inc., 2012), various synchronization intervals were extracted as the temporal distance of a potential synchronization point from the metronome pulse (Fig. 1). Therefore, if such a synchronization point coincided precisely with the metronome, the corresponding synchronization interval was 0; if the synchronization point was located after the metronome beat, it yielded positive values (grey bars below the spectrogram in Fig. 1), whereas location before the metronome beat produced negative values of the synchronization interval (dark bars). In initial analyses, all segment boundaries were considered as potential synchronization points, and two additional moments within the word were chosen as well: points within the first syllable were located which corresponded to (a) the energy function maximum (maxE) and (b) the difference function maximum (maxD).

With regard to maxE, the energy values were calculated from 40-ms segments windowed with a rectangular window (since we do not focus on particular spectral properties) with a 44-sample shift (yielding approximately a 1-ms shift given the sample rate of 32 kHz). These values were then filtered via a moving-average filter of order 6 to obtain a smoothed energy function. The time axis of this function was computed as temporal midpoints of the original 40-ms segments with a correction for the group delay of the filter (which corresponds to order/2 = 3 segments). The difference function – used for establishing the maxD point – was evaluated from the previously smoothed energy function as the difference of the actual and previous sample, and the result was smoothed via moving-average filter of order 10. The time axis values were also corrected with respect to the group delay of the filter (in this case, 5 segments). In order to avoid the initial highly variable random values due to fricative sounds in some words, the maximum of smoothed energy difference function was searched starting from the first segment of the original signal with zero-crossing rate less than 4000, which was calculated according to Eq. (1):

$$zcr = \frac{f_s}{2T}\sum_{t=1}^{T-1} \mathbf{I}\{x_t x_{t-1} < 0\} \tag{1}$$

where T is the number of samples in a segment, x_t is a sample of the segment at a discrete time, f_s is the sampling rate, and the indicator function $\mathbf{I}\{xtxt-1<0\}$ yields 1 if its argument is true and 0 otherwise.

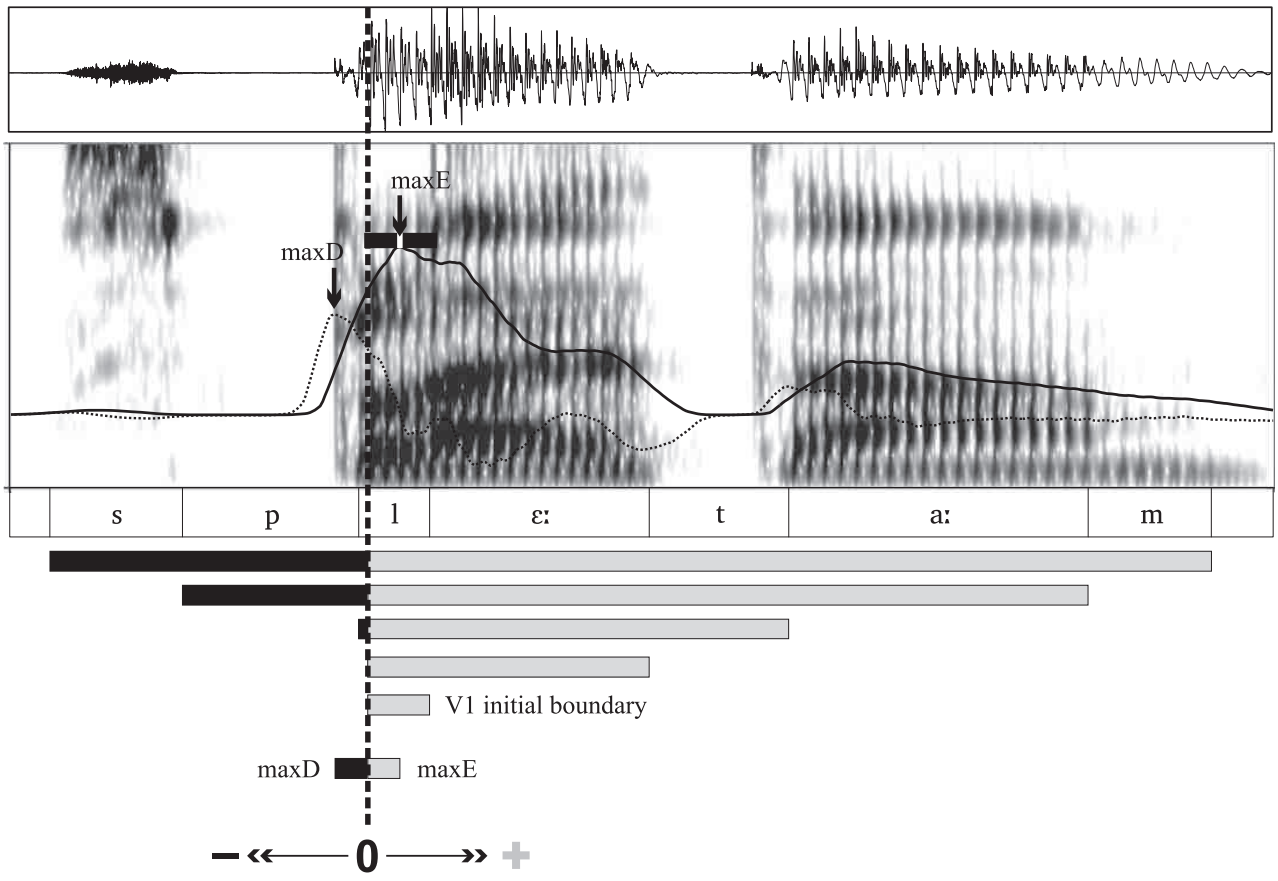

Fig. 1. Different synchronization intervals, i.e., time intervals between a potential synchronization point – a segment boundary, maxE, or maxD – and the metronome pulse. Bars below the spectrogram represent the size and polarity of the intervals. Energy envelope used for the calculation of the maxE point is indicated by a solid curve (black bar above it stands for the relevant window). The difference function used for the calculation of the maxD point is indicated by a dotted curve. Position of the metronome pulse in this example token is marked by a dashed vertical line.

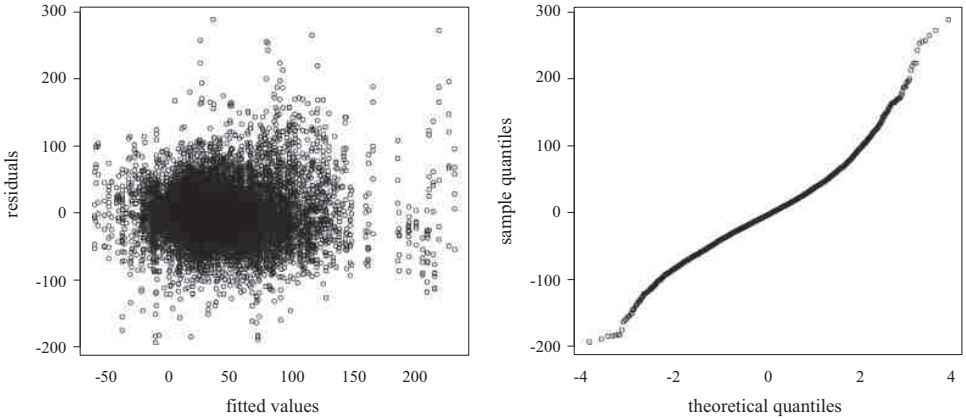

Fig. 2. Residuals plotted against fitted values of the model comprising fixed effects of tempo, initial onset and final coda (left) and a *Q–Q* plot checking normality of the distribution of residuals (right).

2.5. Statistical analyses

Statistical analyses were performed in the R software (R Core Team, 2015) and the R package *lme*4 (Bates, Maechler, Bolker, & Walker, 2015). Instead of using by-subject and by-item analyses of variance, we employed linear mixed-effects (LME) modelling to resolve the non-independence of observations from the same speaker or to the same item by adding a random effect for both subject and word. Each subject/word is thus assumed to have a different baseline value (random intercept). In addition, we included random slopes to test for interaction between these random effects and the fixed effects (TEMPO: slow, fast; INITIAL ONSET COMPLEXITY: C, CC, CCC; WORD-FINAL CODA: none, obstruent, sonorant). The significance of individual effects or interactions was tested by comparing a full

model (comprising the factor or interaction in question) to a reduced model in which the factor/interaction was excluded; the evaluation used standard likelihood ratio tests.

Several steps were taken to ascertain that our data were suitable for LME modelling. Residuals were plotted against fitted values to check whether the model met the assumption of linearity and homoscedasticity (Fig. 2 on the left). In this example, there was no nonlinear relationship, and the residuals were approximately equal across the range of the fitted values. Furthermore, the *Q*–*Q* plot in Fig. 2 on the right reveals that the residuals were in essence normally distributed. All models reported further yielded similar characteristics.

3. Results

3.1. Points of synchronization and synchronization intervals

The analyses in this section are based on *sextets* ($n=1768$; 24 speakers $\times$ 2 tempi $\times$ 37 targets minus eight discarded trials), as the six repetitions of a target word within one experimental trial were averaged, yielding a mean and a standard deviation value for each synchronization parameter. The standard deviation (SD, henceforth *sextet SD*) was assumed to reflect the inconsistency of speakers within a trial and, ultimately, their ability to lock the articulation to the periodic metronome pulses.

The p-centre is often hypothesized to be located near the initial boundary of the stressed vowel (i.e., its acoustic onset) or within the consonantal part that precedes it. We therefore compared several synchronization points in this range (C3/C2/C1 initial boundaries, V1 initial boundary, energy maximum and difference function maximum) as determiners of mean synchronization intervals, summarized in Table 1. The initial boundaries of all consonants constituting the syllable onset were located before the metronome pulse in most cases (negative values), while the stressed vowel boundary occurred on average 45 ms after the metronome (39 ms and 51 ms in slow and fast tempo, respectively). In other words, the metronome pulse, possibly coinciding with the p-centre, generally lay somewhere within the syllable onset (ahead of the vowel itself). However, an even closer approximation than the V1 initial boundary seemed to be on the one hand the moment of difference function maximum (maxD), which was located only 26 ms *after* the assumed p-centre in the slow tempo (40 ms in the fast tempo), and on the other hand the C1 initial boundary, positioned 39 ms and 20 ms *before* the metronome pulse (for slow and fast tempo, respectively). Importantly, although the C1 boundary was closer to the metronome pulse in comparison to the V1 boundary, the boundary of the vowel was associated with a lower variation (especially in the items analysis), reflected in the SD. It is also evident from the table that in the fast tempo condition the speech production shifted so that the metronome pulse occurred more ahead of the vowel (i.e., closer to the beginning of the word).

The relations can be captured in synchronization profiles (for individual words or speakers) such as the one in Fig. 3. The *x* axis represents the articulation of segments, while the *y* axis shows the synchronization interval for each synchronization point (a segment boundary or maxE/maxD). The lines interpolate between the points at $\pm$ 1.96SD. The metronome pulse (approximation of the given word's p-centre) is shown by the dashed horizontal line at time zero. In this particular word (/ˈslɛːtaːm/), the metronome pulse was quite close to the initial boundary of [l], especially in the slow tempo. The maxD point was also a good approximation, while the initial boundary of [ɛː] was further away, as was the maxE point. The rest of the word naturally diverged even more. A turning point occurred for the two tempi in the intervocalic consonant, as the first syllable had lower synchronization intervals in the slow tempo than in the fast tempo (relative to metronome all boundaries started earlier in the slow tempo), whereas the second syllable presented the opposite (all boundaries started later in the slow tempo). This is to be expected since the faster tempo is linked with shorter segment durations in general.

The question of which synchronization point to use in further analyses can be resolved by considering the stability of the individual points. In addition to the variability of the mean synchronization intervals reported in Table 1, Table 2 displays mean sextet SDs, reflecting the consistency of speakers within individual trials. Although the differences are not marked, there was a tendency for the V1 initial boundary to have a more consistent synchronization with the metronome within individual sextets than other boundaries or the moments of maxE and maxD. Also, the standard deviation of the mean sextet SDs was again lowest for V1.

Table 1
Mean synchronization intervals derived for different synchronization points within the word (V1 = stressed vowel; C1–C3 = first to third consonant in the initial onset, counted from the closest position to the stressed vowel, i.e., (C3)(C2)C1V1; maxE = moment of energy maximum; maxD = moment of difference function maximum), separately for slow and fast tempo. The data are based on sextets, i.e. after averaging across repetitions. Two sets of means and SDs are given: subscript i designates analysis by items (averaging across participants; *N* varies for different synchronization points since C2 and C3 were present only in a subset of target words) and subscript s analysis by subjects (averaging across words; *N* is thus always 24).

Synchronization point	Synchronization interval (ms)									
	Slow tempo					*Fast tempo*				
	$Mean_{i,\ s}$	N_i	SD_i	N_s	SD_s	$Mean_{i,\ s}$	N_i	SD_i	N_s	SD_s
C3 initial boundary	−146	3	14.8	24	52.5	−99	3	26.0	24	53.3
C2 initial boundary	−114	22	36.4	24	34.7	−79	22	33.8	24	33.4
C1 initial boundary	−39	37	44.4	24	28.5	−20	37	46.0	24	27.7
V1 initial boundary	39	37	25.1	24	24.6	51	37	28.2	24	25.9
maxE	68	37	27.9	24	27.4	78	37	28.8	24	26.6
maxD	26	37	26.1	24	26.7	40	37	26.5	24	26.9

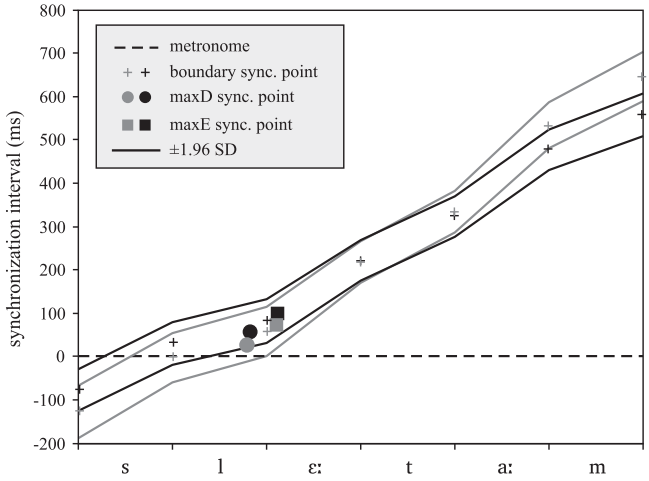

Fig. 3. Mean synchronization intervals for /ˈslɛːtaːm/ derived for different synchronization points within the word (all segment boundaries; maxE = moment of energy maximum; maxD = moment of derivation function maximum). Separately for slow (n = 24; in grey) and fast tempo (n = 24; in black).

Table 2
Mean sextet standard deviations derived for different potential synchronization points within the word (V1 = stressed vowel; C1–C3 = first to third consonant in the initial onset, counted from the closest position to the stressed vowel, i.e., (C3)(C2)C1V1; maxE = moment of energy maximum; maxD = moment of difference function maximum), separately for slow and fast tempo. The data are based on sextets, i.e. after averaging across repetitions. Subscript i designates analysis by items (averaging across participants; *N* varies for different synchronization points since C2 and C3 were present only in a subset of target words) and subscript s analysis by subjects (averaging across words; *N* is thus always 24).

Synchronization point	Sextet standard deviation (ms)									
	Slow tempo					*Fast tempo*				
	$Mean_{i,\ s}$	N_i	SD_i	N_s	SD_s	$Mean_{i,\ s}$	N_i	SD_i	N_s	SD_s
C3 initial boundary	34	3	4.1	24	11.4	28	3	5.1	24	14.8
C2 initial boundary	34	22	4.5	24	8.1	26	22	3.8	24	6.1
C1 initial boundary	32	37	3.3	24	6.3	25	37	4.7	24	5.1
V1 initial boundary	30	37	3.1	24	5.8	24	37	3.8	24	4.8
maxE	35	37	5.1	24	6.6	27	37	5.8	24	5.4
maxD	33	37	4.1	24	6.3	26	37	4.6	24	5.6

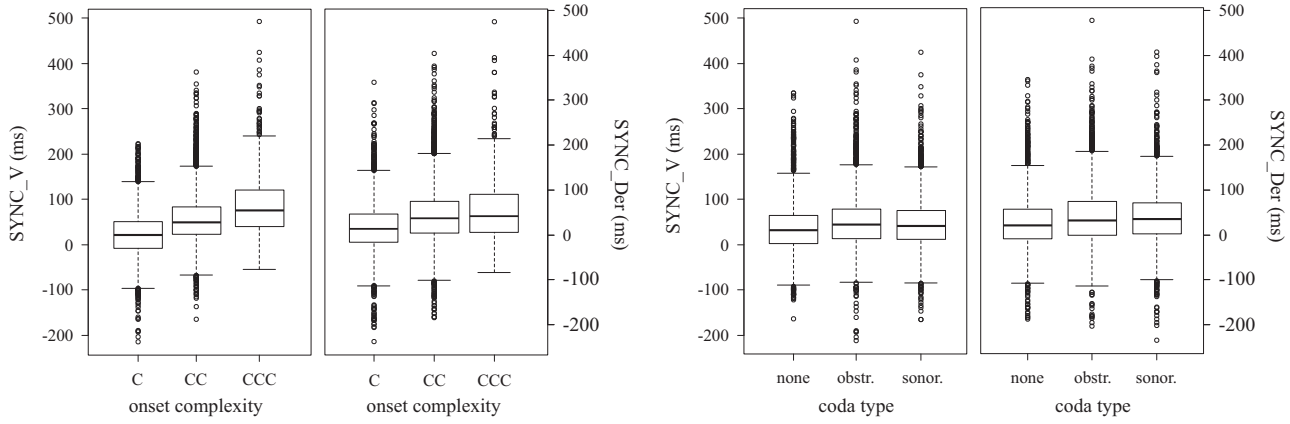

Fig. 4. The effect of onset complexity (left) and coda type (right) on synchronization intervals (SYNC_V and SYNC_Der; left and right within the factors, respectively). Boxes indicate quartile ranges, whiskers denote ± 1.5 IQR from the quartiles, and outliers are plotted as individual points.

In sum, our final decision was motivated primarily by theoretical grounds – relevance of the vowel and of energy distribution reported in the literature – but it was also supported by the analysis of (1) the distance of candidate synchronization points from the metronome, (2) the variability of this distance, (3) mean synchronization inconsistency of speakers within sextets with regard to individual candidate points and (4) the variability of this parameter. Therefore, analyses reported below will be restricted to two parameters only, SYNC_V (synchronization interval based on the first vowel initial boundary) and SYNC_Der (synchronization interval based on the moment of difference function maximum, maxD, corresponding to the fastest increase of energy). Since sonorants, compared to obstruents, tend to attract the maxD point to themselves, the two parameters are expected to differ most in words with prevocalic sonorant consonants (/l/, /r/).

In light of the ambiguity regarding p-centres and metronome pulses (see Section 4), it should be noted that the first and foremost merit of our analyses is not in giving absolute values in milliseconds but rather in showing relations between various words or conditions.

3.2. Effect of phonotactic structure and tempo on synchronization intervals

A set of LME models was constructed from the full data and evaluated by likelihood ratio tests (separately for SYNC_V and SYNC_Der as the dependent measure). In addition to the random effects, a basic model included the fixed effects of onset complexity, type of coda and tempo. Generally, it is apparent from Fig. 4 that the energy-related measure SYNC_Der yields similar results as SYNC_V, although there are some differences between them. Onset affected SYNC_V ($\chi^2(2)=31.67$, $p<0.001$), increasing it by approximately 32 ms (± 5.2 SEs) for CC onsets and by 66 ms (± 10.3 SEs) for CCC onsets. SYNC_Der was also significantly influenced by onset ($\chi^2(2)=14.33$, $p<0.001$), with an increase of 25 ms (± 6.8 SEs) and 41 ms (± 10.6 SEs) for CC and CCC onsets, respectively. However, although sonorant and obstruent codas increased the synchronization intervals in comparison to the no coda condition, the effect of coda type did not prove to be significant ($\chi^2(2)=5.71$, $p>0.05$ for SYNC_V, $\chi^2(2)=3.66$, $p>0.05$ for SYNC_Der). These two effects are captured in Fig. 4 showing that with more complex syllable onsets the speech activity was shifted forward relative to the metronome pulse (the presumed p-centre of the word). Adding a final obstruent to the word had a similar, but substantially weaker influence.

A comparison of models with and without the interaction between the two fixed effects showed that there was no significant interaction ($\chi^2(4)=0.50$, $p>0.05$ for SYNC_V, $\chi^2(4)=1.13$, $p>0.05$ for SYNC_Der). Potentially, a subset of 18 words with a balanced phonotactic structure (Table 3) might capture the influence of syllable onsets and word-final codas (and their interaction) more directly. A new series of models was constructed, separately for SYNC_V and SYNC_Der, with the same random and fixed effects as previously but based on this reduced dataset. The relations between the factors were very similar to the full data reported above. Most importantly, there was no interaction between onset complexity and type of coda.

Furthermore, in the full dataset, tempo affected SYNC_V and SYNC_Der significantly ($\chi^2(1)=8.46$, $p<0.01$ and $\chi^2(1)=10.42$, $p<0.01$, respectively). As Table 1 already suggested, words were articulated later in the fast tempo relative to metronome pulses, the synchronization interval being larger in the fast tempo by 11 ms (± 3.9 SEs) for SYNC_V and by 14 ms (± 4.0 SEs) for SYNC_Der. Although tempo did not interact with onset complexity ($\chi^2(2)=0.64$, $p>0.05$ and $\chi^2(2)=2.50$, $p>0.05$ for the two measures), the interaction with type of coda was significant ($\chi^2(2)=8.60$, $p<0.05$ for SYNC_V and $\chi^2(2)=6.31$, $p<0.05$ for SYNC_Der). Specifically, when the interaction is taken into account in the model, the influence of coda type has different characteristics in the slow vs. fast tempo. For SYNC_V, an obstruent in the coda was associated with an increase in synchronization interval of 18 ms in the slow tempo and 11 ms in the fast tempo, whereas a sonorant in the coda led to an increase of only 8 ms in the slow tempo and 15 ms in the fast tempo (SEs ranged between 4 and 5 ms). Thus, the effect of tempo on coda obstruents was opposite to that found for sonorant codas. For SYNC_Der, the relations were largely analogical.

However, the effect of tempo seems to be massively influenced by 3 out of 24 speakers who demonstrate a disproportionate increase (over 40 ms) in the synchronization intervals for the fast tempo; if these speakers are excluded from analyses, the difference drops down to only 6 ms (SYNC_V) and 8 ms (SYNC_Der), which is nevertheless still significant, albeit with lower test criteria ($\chi^2(1)=4.92$, $p<0.05$ for SYNC_V; $\chi^2(1)=8.03$, $p<0.01$ for SYNC_Der).

Individual words also differed in the amount of change they allowed between the two tempi. It appears that vowel length was the most significant factor, since words of the type CV or CCV formed a group with a slow-fast difference of 5 ms (for both parameters), whereas CVː, CCVː and CCCVː formed another group with a slow-fast difference of 16 ms and 20 ms (for SYNC_V and SYNC_Der, respectively). The distinct behaviour of the two groups is maintained even after excluding the three speakers from analyses; however, the slow-fast difference in the former group decreases to 0 ms (for both parameters) and in the latter group to 12 ms and 16 ms (for SYNC_V and SYNC_Der, respectively).

3.3. Exploration of individual words

Appendix includes synchronization intervals for individual target words, ordered by the phonological complexity of the first syllable (onset complexity and vowel length) and the size of SYNC_V. Fig. 5 is a succinct expression of the fact that the metronome pulse was located increasingly further ahead of the vowel as more complex structures were synchronized with the metronome. SYNC_V

Table 3
Subset of target words according to initial onset complexity and word-final coda type.

Initial onset	Final coda		
	No coda	/m/	/ʃ/
C	cɛkaː	cɛkaːm	cɛkaːʃ
C	lɛːtaː	lɛːtaːm	lɛːtaːʃ
CC	stɛːkaː	stɛːkaːm	stɛːkaːʃ
CC	slɛːtaː	slɛːtaːm	slɛːtaːʃ
CC	ʃcɛkaː	ʃcɛkaːm	ʃcɛkaːʃ
CCC	splɛːtaː	splɛːtaːm	splɛːtaːʃ

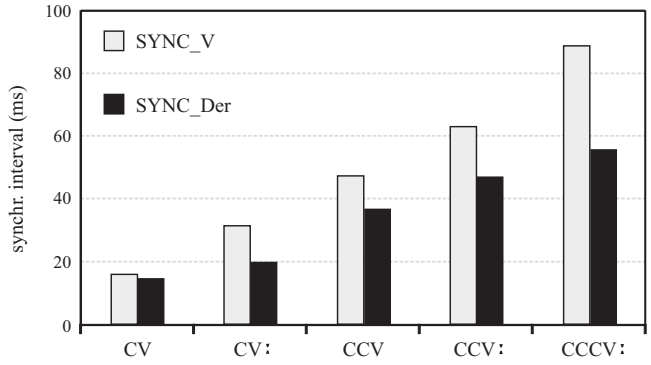

Fig. 5. Synchronization intervals for sets of words with increasing phonological complexity of the first syllable. Individual words are provided in the Appendix. SYNC_V is related to the initial boundary of the first vowel, while SYNC_Der to maxD (the moment of fastest increase in energy).

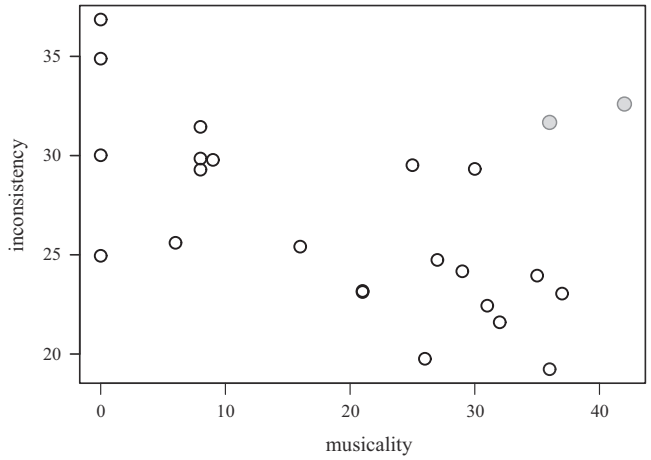

Fig. 6. A scatterplot of synchronization inconsistency scores and musicality scores for individual speakers (24 speakers in full data set; 2 deviating respondents indicated with grey circles). Inconsistency scores were computed as the mean (over 74 test items) of standard deviations (over 6 repetitions of the target word within an item) of the distances between the metronome and the initial boundary of the first vowel (i.e., SYNC_V). Musicality scores comprise Years, Motivation and Ensemble Play. For correlation coefficients see text.

and SYNC_Der were comparable only in CV words, and differed widely in the other conditions. However, the situation is not independent of the phonetic structure of the groups: it happens to be the case that the CV set was composed only of obstruents, whereas the CV: set included four /l/-initial words and only two with an initial obstruent; further, the CCV and CCV: sets were balanced for obstruents vs. sonorants preceding the vowel, and the CCCV: set comprised only /spl/. This is a disadvantage of using real words, rather than constructed non-words as others have generally employed in their studies. The differences between the two parameters might therefore be explained in terms of the prevocalic segment: SYNC_Der is smaller than SYNC_V in words with /l/ or /r/ preceding the vowel (i.e., the maxD point is 27 ms closer to the metronome), but not in words with obstruents in the prevocalic position (the maxD point approximates the beginning of the vowel). For instance, /st/-initial words had comparable SYNC_V and SYNC_Der values (55 ms and 53 ms, respectively), but /sl/-initial words yielded a SYNC_V of 65 ms and a SYNC_Der of 37 ms. The moment of difference function maximum is thus closer to the supposed p-centre, approximated by the metronome, than the beginning of the vowel is in these words.

A more detailed look at the words reveals that in the CV set, corresponding to the simplest structures, the p-centre is influenced by the phonetic segments in the onset: words with plosives had significantly lower synchronization intervals than words with fricatives ($t(427) = -7.54$, $p < 0.001$ for SYNC_V; $t(427) = -9.62$, $p < 0.001$ for SYNC_Der). However, there were no clear trends in the more complex structures, apart from the fact that the affricate /t͡s/ seemed to belong to a set with a larger onset size, as if its two articulatory parts formed two segments: /t͡sɛ:zar/ and /st͡sɛ:na:r̥/ behaved as CC and CCC words, respectively. This was also supported by onset duration ([spl] = 211 ms, [st͡s] = 193 ms, other CC onsets = 160 ms).

3.4. Synchronization skills and musicality

We computed two parameters that were expected to describe the performance of individual subjects. The **synchronization inconsistency score** was based on the inconsistency in synchronization within sextets (expressed as the Standard Deviation of a sextet for the SYNC_V interval). For each subject, the score was determined as the mean of these sextet SDs, i.e., averaged over all target words (37 words × 2 tempi). The second evaluation of the subjects was external to the experiment, consisting of self-reported information about their musical background (see Section 2.2). The answers were converted to points, which were then added up into a **musicality score**. It should be noted that a higher value of the synchronization inconsistency score (greater sextet SD) describes a

worse performance of the subject, while a higher value of the musicality score denotes better musical aptitude and background. Thus, a negative correlation between the two parameters was predicted.

The results are plotted in Fig. 6, with musicality scores of the speakers on the *x* axis and their inconsistency on the *y* axis. It is obvious that, indeed, the values are not random and there is a moderate trend to achieve lower inconsistencies with higher musicality scores. It appears that our sample includes only two divergent speakers (marked by grey circles in the figure) that could potentially be considered for exclusion from analyses.

Various means of determining the musicality scores had been considered, attaching different weights to the information in the questionnaires, with the aim to find which musical attribute was most closely associated with the performance in the synchronization task. It should be stressed that the correlation analyses were used simply as a means to achieve this goal; the main objective was *not* to prove that there is any correlation between musicality and synchronization inconsistency (i.e., when the musicality score is constructed so as to correlate highly, it would be pointless to claim that the musicality measure correlates with synchronization). Several examples are shown in Table 4 (synchronization scores computed from SYNC_V and SYNC_Der are compared). It is clear that some musical attributes do not provide much opportunity to explain the synchronization skills of the speakers (e.g. musical education, intensity of practice, dancing), while others seem to have higher correlations and, we believe, greater explanatory value as well (years of activity, motivation for musical activities, ensemble play). Therefore, the final musicality score shown in Fig. 6 was computed as a sum of the three attributes with highest individual correlations. Pearson's product–moment correlation between the synchronization inconsistency score and this musicality score reached statistical significance and showed a strong negative relationship. However, the relationship was even stronger when the two outlying speakers (see Fig. 6) were excluded. These speakers reported highest musicianship, but performed poorly in the metronome synchronization task. In any case, the correlation results indicate that subjects with more musical experience – in terms of the given attributes – generally performed better in the synchronization task than those with no or little musical experience.

4. Discussion

To assess the precision of rhythmic behaviour a relevant moment of motion needs to be established as the phase anchor relative to which the location of a p-centre can be compared. A vowel acoustic onset is often believed to be such a point (Barbosa et al., 2005; Fox and Lehiste, 1987a; Marcus, 1981), although other studies may measure the p-centre position relative to the acoustic onset of the syllable (e.g. Chow et al., 2015; Fowler, 1979; Pompino-Marschall, 1989). Our data showed that it is indeed the initial boundary of the stressed vowel that appears to be most stable in the attempts of the subjects to synchronize their word production with metronome pulses (see Table 1 and especially Table 2). However, this fact neither refutes nor supports the hypothesis that the underlying p-centre coincides with the beginning of the vowel and that the surface diversity reflects perturbations caused by the environment; rather, it confirms as correct the search for p-centres in the vicinity of the vowel beginning.

As the survey of literature in the Introduction has shown, the p-centre does not appear to be perfectly aligned with a specific acoustic event (such as the vowel onset, the syllable onset), but is generally believed to occur somewhere within the consonantal segment(s) preceding the stressed vowel. The results from our experiment support these findings: the SYNC_V synchronization intervals (distance of V1 initial boundary from metronome pulse) were predominantly positive, with a mean value of 45 ms (averaged across tempi), while the consonantal synchronization intervals were negative (see Table 1 for specific values). The p-centre thus seems to be closest to the beginning of the prevocalic consonant if more consonants are in the onset. However, it can be argued that the p-centre need not coincide with a specific boundary since aspects of the whole stimulus affect the p-centre location. It might be associated with some articulatory event, such as the beginning of the vocalic gesture, which naturally precedes its acoustic onset (see Fowler, 1983), or a moment within the word that corresponds to some perceptually or acoustically salient event, such as the moment of energy maximum or, as we hypothesized, the moment of the fastest increase of energy within the consonant-vowel transition, to some extent corresponding to the articulatory moment of fastest jaw opening. In terms of distance from the metronome pulse our results favour the SYNC_Der synchronization interval over the SYNC_V (the former was on average 12 ms closer to the metronome beat than the latter). The two measures yielded similar values in words where the segment preceding the vowel was an

Table 4
Correlations between musicality scores and synchronization inconsistency scores (based on either SYNC_V or SYNC_Der). All subjects (left columns) vs. two subjects excluded (right columns). Correlations indicated with boldface were significant at the level of $\alpha = 0.05$.

Musical attribute	Synchronization inconsistency score (SYNC_V)		Synchronization inconsistency score (SYNC_Der)	
	24 speakers	*22 speakers*	*24 speakers*	*22 speakers*
Education	−0.10	−0.21	−0.13	−0.24
Years of Activity	−0.40	**−0.69**	−0.37	**−0.62**
Current Activity + Motivation	−0.33	**−0.49**	−0.38	**−0.52**
Ensemble Play (Entrainment)	**−0.58**	**−0.78**	**−0.54**	**−0.70**
Practice	−0.14	−0.16	−0.19	−0.21
Dance	−0.01	−0.03	−0.07	−0.06
Years + Motivation + Ensemble Play	**−0.47**	**−0.72**	**−0.46**	**−0.67**
Aggregate of all information	−0.39	**−0.61**	−0.40	**−0.59**

obstruent, but differed in words with prevocalic sonorant segments. The main advantage of SYNC_Der was thus associated with targets like /slɛːtaː/ or /frɛːza/, as compared to e.g. /stɛːkaː/.

The current experiment is based on the assumption that the metronome pulse coincides temporally with the position of the p-centre in the given word. However, this is not necessarily the case. As pointed out by one of the reviewers, it is common for participants in sensorimotor synchronization experiments like tapping to anticipate the metronome pulse and tap ahead of time. If there is indeed such an anticipation effect, then the temporal point at which the metronome pulse occurs in the speakers' production may not correspond to the actual p-centre. This could be considered a limitation to our study, but until a sufficiently precise model is provided for the anticipation effect in speech, we can only admit its existence and take it into account while interpreting results. Given the design of the current experiment, the issue cannot be disentangled.

It has been copiously demonstrated that the p-centre changes its position (measured relative to the vowel beginning) in proportion to the duration of the syllable onset (already established for instance by Rapp, 1971 or Allen, 1972); in other words, with more consonants in the syllable onset, the p-centre is located further away from the vowel. Our experiment successfully replicated the results on Czech with a variety of consonantal sequences: C, CC and CCC initial words yielded increasingly larger synchronization intervals. Similarly, /t͡s/ seemed to pattern with the more complex onsets, revealing its ambivalent segmental status as regards its articulatory complexity. Finally, in the simplest structures (with a single C in the syllable onset), the phonetic nature of the segment was of importance. Specifically, in words with initial plosives (both voiced and voiceless) the p-centre – i.e., the metronome pulse – coincided fairly well with the acoustic vowel onset, whereas fricatives attracted the p-centre to themselves. It thus appears that all segments do not behave alike and that the amount of energy, or the changing spectral content, in the prevocalic portion cannot be ignored. An experiment designed specifically to investigate this effect would be needed to validate the results based on the current data. An important conclusion, however, should be drawn anyway. Given that the difference in coda sonorant versus obstruent was in significant interaction in the two tempi, the future rhythm research should be more cautious about collapsing all consonants under the label of "C". Various classes of consonants clearly contribute to the rhythm configurations in their own way.

In contrast to the results regarding initial consonants, our experiment contradicts the expectations based on previous research examining the influence of the rest of the word (i.e., the stressed vowel and the second, unstressed syllable). Several studies on English monosyllables reported an effect of vowel duration, with longer vowels attracting the p-centre to themselves, that is, shifting the p-centre in the opposite direction than longer onsets (Fox and Lehiste, 1987a; Marcus, 1981). However, our data do not seem to support these findings and indicate quite the opposite, as structures with phonologically long vowels had larger SYNC_V (and thus earlier p-centres) than words with short vowels. Moreover, the same effect was associated with words with a final coda consonant as opposed to words ending in a vowel: everything else being equal, the former group demonstrated slightly larger SYNC_V (but not significantly different). Again, this finding contradicts previous results on English, where Marcus (1981) and Cooper et al. (1988) showed that coda consonants added to monosyllables had the opposite effect on p-centre location, and Fox and Lehiste (1987b) found a similar effect of adding an unstressed suffix to a monosyllable. The current results on the Czech language point instead to an effect of complexity: in more complex structures (i.e., with larger onsets and/or with phonologically long vowels and/or with final codas) the p-centre occurs earlier in the word compared to more simple structures. However, it must be noted that the effect of vowel length and the presence of word-final coda is much lower than the effect of onset complexity. Indeed, it is the onset structure that represents the most robust factor in our analyses.

This finding actually offers more perspectives. It may be the case that p-centres, which have been investigated predominantly on Germanic languages, do not exhibit universal characteristics. An indication of this might be the recent study by Chow et al. (2015) examining speech-metronome synchronization in Cantonese, where the influence of coda consonants (in monosyllables) was likewise insignificant. An alternative explanation would be that the unexpected results in Czech are due to the differences in methods. The results of our experiment are not directly comparable to any study since we used (1) natural words as opposed to nonsense syllables, (2) disyllables as opposed to monosyllables, (3) a language with phonologically distinctive vowel length, (4) speech-metronome synchronization, also including two different metronome rates, (5) a sample of 24 speakers. Any of these could influence the results, especially if the effects found in other studies were less robust, such as the coda consonant or vowel duration effects.

Furthermore, although we did not predict any differences in synchronization intervals between the two metronome rates, the faster tempo was associated with larger SYNC_V (and SYNC_Der). The difference decreased to a half when three outlying speakers were excluded, but still remained significant. It is possible that the task was more difficult in the fast tempo so that the subjects might tend to fall behind in their synchronized articulation compared to the slow condition. This seemed to be the case especially with more complex words, such as /stɛːkaːm/, but there were frequent exceptions from both sides (/splɛːtaː/ yielded almost no difference, while /dɛka/ yielded a medium-sized difference). However, in other respects, the speakers behaved similarly in the two tempi, producing identical or highly similar patterns of synchronization. In other words, for the rate of approximately 2–3 syllables per second and 3-4 syllables per second we can expect rather proportionate trends.

The only exception was the effect of word-final codas on the synchronization intervals, as in the fast condition items with /m/ codas were treated by the speakers similarly to items with /ʃ/ codas, but in the slow condition /m/ codas patterned rather with words that ended in a vowel. However, although the interaction of tempo with coda type was significant in the LME model ($p<0.05$), care must be taken in interpretation because the standard errors were quite substantial given the small differences between conditions. Nevertheless, this finding is of great interest and, if generalizable to other segments, may have important implications for investigating rhythm in general. If a sonorant consonant added to a vowel does not change the synchronization pattern of the word, while an obstruent does, it might be advisable not to separate the sonorant from the vowel when computing global rhythm metrics (%V, ΔC, PVI). Therefore, we plan to follow these tentative results with a more suitable experiment in the future to investigate the issue further.

Finally, the present study was quite exceptional in the number of subjects that participated in the experiment, resonating with the objections in Kohler (2009) that the human population displays substantial variance in the "rhythm aptitude" and any research should take this fact into account. The only other large-scale study we are aware of is Chow et al. (2015) who, unfortunately, do not provide any information about individual subjects. The 24 speakers we analysed differed in their ability to synchronize speech with the metronome (assumed to be reflected in the standard deviations of individual trials, or "sextets") but the differences were not marked. Interestingly, the synchronization scores correlated significantly with the musicality scores derived from questionnaires about the subjects' musical background: lower scores on the musicality scale were associated with greater inconsistency in synchronization (higher mean sextet SDs). This stands in contrast with the results of Chow et al. (2015) who did not find any effect of musical background. However, they measured it as the number of years of musical training, whereas our analysis included other related issues as well, such as motivation and perhaps most importantly the experience with ensemble (collective) playing (see Table 4).

5. Conclusions

The present speech-metronome synchronization study based on Czech disyllabic words was aimed to replicate findings about the location of p-centres that have been established on a small sample of languages and, in many cases, on stimuli of limited complexity. In agreement with other studies (e.g. Fowler, 1979; Marcus, 1981; Morton et al., 1976), the most important factor that contributed to the perception of a rhythmical beat was found to be the complexity of the syllabic onset: the position of the p-centre was found to be further ahead from the centre of the vowel when more consonants were included in the onset. The effect was very robust and identical in two metronome rates. Vowel length and final coda consonants exerted a much smaller influence, although the effect was analogical: more complex structures were associated with the p-centre more ahead of the stressed vowel. The direction of this shift contrasts with earlier findings of Marcus (1981), Fox and Lehiste (1987a, b), and Cooper et al. (1988). Furthermore, the moment of fastest increase of energy appeared to be generally a better synchronization point (closer to the metronome pulses) than the initial boundary of the stressed vowel. Lastly, the ability of speakers to synchronize their articulations with an isochronous auditory sequence was significantly correlated with musicality scores based on their self-reported musical background.

Acknowledgements

The research was supported by the Charles University Grant Agency (GAUK) under Grant 834213, and by the Charles University in Prague programme for science development P10-Linguistics. The authors would like to thank Eliška Churaňová and Tomáš Bořil for their important help with the project.

Appendix

List of target words ordered by phonological complexity of the first syllable and SYNC_V. In the verbs, the suffixes *-m*, -š and *-Ø* designate, respectively, 1st, 2nd and 3rd person singular present tense.

Word	Translation	First syllable	SYNC_V (ms)	SYNC_Der (ms)
ɟɛkan	dean	CV	−3	−5
cɛkaː	he/she is flitting	CV	−1	−4
pɛkaɾ̝	baker	CV	2	−1
cɛkaːm	I am flitting	CV	8	5
dɛka	blanket	CV	10	0
cɛkaːʃ	you are flitting	CV	20	16
ʃɛdaː	grey	CV	25	27
sɛdaːm	I am sitting down	CV	36	40
sɛnaːt	senate	CV	45	52
tɛːma	theme	CVV	18	35
lɛːtaː	he/she is flying	CVV	25	1
lɛːtaːm	I am flying	CVV	29	7
lɛːkaɾ̝	doctor	CVV	32	11
lɛːtaːʃ	you are flying	CVV	41	20
t͡sɛːzar	emperor, Caesar	CVV	43	44
ʃcɛkaː	he/she is barking	CCV	32	28
plɛsat	to rejoice	CCV	38	11
ʃcɛkaːm	I am barking	CCV	39	33
plɛna	nappy	CCV	41	17
ʃcɛpaːn	Steven	CCV	45	43
slɛzan	a Silesian	CCV	56	33
ʃpɛnaːt	spinach	CCV	60	63

ʃcɛkaːʃ	you are barking	CCV	68	64
frɛːza	milling machine	CCVV	47	22
mlɛːkar̥	milkman	CCVV	51	−1
stɛːkaːʃ	you are flowing down	CCVV	52	49
stɛːkaːm	I am flowing down	CCVV	53	50
stɛːkaː	he/she is flowing down	CCVV	60	58
slɛːtaː	he/she is flying down	CCVV	63	33
slɛːtaːʃ	you are flying down	CCVV	63	37
krɛːcan	a Cretan	CCVV	66	47
stɛːnaːm	I am moaning	CCVV	67	73
slɛːtaːm	I am flying down	CCVV	69	42
st͡sɛːnaːr̥	screenplay	CCVV	100	103
splɛːtaː	he/she is interweaving	CCCVV	78	44
splɛːtaːʃ	you are interweaving	CCCVV	91	60
splɛːtaːm	I am interweaving	CCCVV	97	64

References

Allen, G. C. (1972). The location of rhythmic stress beats in English (I & II). *Language and Speech*, *15* 72–100 and 179–195.

Barbosa, P. A., Arantes, P., Meireles, A. R., & Vieira, J. M. (2005). Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors. In: *Proceedings of interspeech 2005* (pp. 1441–1444). Lisbon: ISCA.

Bartoň, T., Cvrček, V., Čermák, F., Jelínek, T., & Petkevič, V. (2009). *Statistiky češtiny*. Praha: Nakladatelství Lidové noviny.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-8. ⟨http://CRAN.R-project.org/package=lme4⟩.

Boersma, P. & Weenink, D. (2013). *Praat—Doing phonetics by computer (version 5.3.41) [Computer software]*. Retrieved from ⟨http://www.praat.org/⟩.

Buxton, H. (1983). Temporal predictability in the perception of English speech. In A. Cutler, & D. R. Ladd (Eds.), *Prosody: Models and measurements* (pp. 111–121). Berlin: Springer-Verlag.

Chow, I., Belyk, M., Tran, V., & Brown, S. (2015). Syllable synchronization and the P-center in Cantonese. *Journal of Phonetics*, *49*, 55–66.

Churaňová, E. (2013). The consonantal–vocalic structure of the Czech word and stress group. *AUC Philologica 1/2014, Phonetica Pragensia XIII* (pp. 79–90).

Cooper, A. M., Whalen, D. H., & Fowler, C. (1986). P-Centers are unaffected by phonetic categorization. *Perception & Psychophysics*, *39*, 187–196.

Cooper, A. M., Whalen, D. H., & Fowler, C. (1988). The syllable's rhyme affects its p-center as a unit. *Journal of Phonetics*, *16*, 231–241.

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, *26*(2), 145–171.

Cummins, F. (2009). Rhythm as an affordance for the entrainment of movement. *Phonetica*, *66*, 15–28.

Cutler, A., & Otake, T. (2002). Rhythmic categories in spoken-word recognition. *Journal of Memory and Language*, *46*, 296–322.

de Jong, K. (1994). The correlation of P-center adjustments with articulatory and acoustic events. *Perception & Psychophysics*, *56*(4), 447–460.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35/1*, 116–124.

Fowler, C. A. (1979). "Perceptual centers" in speech production and perception. *Perception & Psychophysics*, *25*, 375–388.

Fowler, C. A. (1983). Converging sources of evidence in spoken and perceived rhythms of speech: Cyclic production of vowels in sequences of monosyllabic stress feet. *Journal of Experimental Psychology: General*, *112*, 386–412.

Fowler, C. A., & Tassinary, L. G. (1981). Natural measurement criteria for Speech: The anisochrony illusion. In J. Long, & A. Baddeley (Eds.), *Attention and Performance IX* (pp. 521–535). Hillsdale, N.J: Lawrence Erlbaum.

Fox, R. A. (1987). Perceived p-center location in English and Japanese. In B. D. Joseph, & A. M. Zwicky (Eds.), *A Festschrift for Ilse Lehiste* (pp. 11–20). Columbus, OH: Department of Linguistics of Ohio State University.

Fox, R. A., & Lehiste, I. (1985). The effect of final consonant structure on syllable onset location. *Journal of the Acoustical Society of America*, *77*, S54.

Fox, R. A., & Lehiste, I. (1987a). The effect of vowel quality variations on stress-beat location. *Journal of Phonetics*, *15*, 1–13.

Fox, R. A., & Lehiste, I. (1987b). The effect of unstressed affixes on stress-beat location in speech production and perception. *Perceptual and Motor Skills*, *65*, 35–44.

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1–2), 113–126.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, *31*, 423–445.

Harsin, C. A. (1997). Perceptual-centre modeling is affected by including acoustic rate-of-change modulations. *Perception and Psychophysics*, *59*, 243–251.

Hoequist, C. E. (1983). The perceptual center and rhythm categories. *Language and Speech*, *26*, 367–376.

Howell, P. (1984). An acoustic determinant of perceived and produced anisochrony. In *Proceedings of the 10th ICPhS* (pp. 429–433). Utrecht.

Howell, P. (1988a). Prediction of P-center location from the distribution of energy in the amplitude envelope: I. *Perception and Psychophysics*, *43*, 90–93.

Howell, P. (1988b). Prediction of P-center location from the distribution of energy in the amplitude envelope: II. *Perception and Psychophysics*, *43*, 99.

Kelso, J. A. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology*, *246*, R1000–R1004.

Kelso, J. A. (1995). *Dynamic patterns*. Cambridge, MA: MIT Press.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*, 1208–1222.

Knight, S. & Cross, I. (2012). Rhythms of persuasion: The perception of periodicity in oratory. In: *Proceedings of perspectives on rhythm and timing* (p. 27). Glasgow.

Kohler, K. (2009). Rhythm in speech and language: A new research paradigm. *Phonetica*, *66*, 29–46.

Kohno, M. (1995). Two different systems for rhythm processing and their hierarchical relation. In: *Proceedings of the 13th ICPhS* (pp. 94–97). Stockholm.

Lee, C. S., & Todd, N. P. M. (2004). Towards an auditory account of speech rhythm: Application of a model of the auditory primal sketch to two multilanguage corpora. *Cognition*, *93*, 225–254.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, *5*, 253–263.

Marcus, S. (1981). Acoustic determinants of perceptual centre (p-centre) location. *Perception & Psychophysics*, *30*(3), 247–256.

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centres (P-centres). *Psychological Review*, *83*, 405–408.

Patel, A., Löfquist, A. & Naito, W. (1999). The acoustics and kinematics of regularly timed speech: A database and method for the study of the p-centre problem. In: *Proceedings of the 14th ICPhS* (pp. 405–408). San Francisco.

Pollák, P., Volín, J. & Skarnitzl, R. (2007). HMM-based phonetic segmentation in Praat environment. In: *Proceedings of SPECOM* 2007 (pp. 537–541). Moscow.

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-centre phenomenon. *Journal of Phonetics*, *17*, 175–192.

Port, R. (2003). Meter and speech. *Journal of Phonetics*, *31*, 599–611.

Port, R., Tajima, K. & Cummins, F. (1996). Self-entrainment in animal behavior and human speech. In *Online proceedings of the* 1996 *Midwest Artificial Intelligence and Cognitive Science Conference*. Retrieved from ⟨http://www.cs.indiana.edu/event/maics96/Proceedings/Port/port.html⟩.

Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, *35*, 353–362.

Quené, H., & Port, R. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica*, *62*(1), 1–13.

R Core Team (2015). *R: A language and environment for statistical computing (version 3.2.1) [Computer software]*. Vienna: R Foundation for Statistical Computing. ⟨http://www.R-project.org/⟩.

Rapp, K. (1971). A study of syllable-timing. In: *Speech transmission laboratory: Quarterly progress and status report* 1 (pp. 14–19). Stockholm: Royal Institute of Technology.

Repp, B. H. (2005). Sensorimotor synchronization: A review of the tapping literature. *Psychonomic Bulletin and Review*, *12*(6), 969–992.

Repp, B. H. (2010). Sensorimotor synchronization and perception of timing: Effects of music training and task experience. *Human Movement Science*, *29*, 200–213.
The MathWorks, Inc.. MATLAB Release 2012b [Computer software]. Massachusetts: Natick.
Tuller, B., & Fowler, C. (1980). Some articulatory correlates of perceptual isochrony. *Perception and Psychophysics*, *27*(4), 277–283.
Turvey, M. T. (1990). Coordination. *American Psychologist*, *45*(8), 938–953.
Volín, J. (2010). On the significance of the temporal structuring of speech. In M. Malá, & P. Šaldová (Eds.), *...for thy speech bewrayeth thee (A Festschrift for Libuše Dušková)* (pp. 289–305). Praha: FF UK.
Volín, J., Churaňová, E., & Šturm, P. (2014). P-centre position in natural two-syllable Czech words. In N. Campbell, D. Gibbon, & D. Hirst (Eds.), *Proceedings of the 7th international conference on speech prosody* (pp. 920–924). Dublin: TCD.
Volín, J., Skarnitzl, R., & Pollák, P. (2005). *Confronting HMM-based phone labelling with human evaluation of speech production. Proceedings of interspeech 2005* (pp. 1541–1544)Lisbon: ISCA1541–1544.