Erschienen in: Journal of Phonetics. Elsevier. 2023, 100, 101263
https://doi.org/10.1016/j.wocn.2023.101263

An acoustic study of rhythmic synchronization with natural English speech

Tamara Rathcke [a], Chia-Yuan Lin [b]

[a] *Department of Linguistics, Universität Konstanz, Konstanz, Germany*
[b] *Department of Psychology, University of Huddersfield, Huddersfield, United Kingdom*

A R T I C L E   I N F O

Keywords:
Speech rhythm
Prosody
Sensorimotor synchronization
P-center
Syllable nucleus
Pitch and Periodic Power
Empirical Mode Decomposition

A B S T R A C T

Sensorimotor synchronization as a means of studying rhythmic perception-action coupling has been extensively researched across a large number of temporally regular structures including music while little is known about synchronization with speech. The present study fills this gap by applying a sensorimotor synchronization paradigm to natural speech and studying acoustic landmarks that may serve as perceptual anchors of rhythmic movement in spoken sentences. Five rhythmically relevant types of acoustic landmarks were identified in twenty sentences of English containing syllables with vocalic and non-vocalic nuclei. The landmarks were either manually defined or algorithm-generated and included nucleus onsets, peaks and onsets of inter-syllabic and inter-stress timescales, moments of the fastest energy change (approximating the P-center location), and timepoints of combined pitch and periodic power. Sensorimotor synchronization data from 32 native English participants were examined with regards to the location of an increased synchronization activity in the proximity of the predefined landmarks. The results demonstrated that participants synchronized with syllable-size units regardless of the type of syllable nucleus (vowel or consonant) and that their taps were consistently timed close to nucleus onsets. Hereby, the manually defined nucleus onsets predicted synchronization peaks as well as the algorithm-generated moments of the fastest energy change around nucleus onsets (i.e., a model of the P-center location) did. In contrast, other landmarks did not constitute a stable acoustic anchor of sensorimotor synchronization with English speech. The synchronization performance was not influenced by either acoustic F0-information or by phonological tune specifications. These findings provide new evidence for the proposals that rhythmic attention in natural speech may be locked on to fast spectral changes within a syllable as the smallest structuring unit of prosodic hierarchy.

1. Introduction

1.1. Background

The nature and even the existence of speech rhythm has been a matter of controversial debates for as long as the history of the term reaches (Kohler, 2009; Nolan & Jeon, 2014; White, Mattys, & Wiget, 2012). In light of the controversy, researchers have been reluctant to even attempt a definition of rhythm in language (Cummins, 2012). Early – much criticized – definitions tended to equalize language rhythm with isochrony, i.e., equal temporal spacing of units, evident either in speech acoustics (Abercrombie, 1967, Ladefoged, 1975) or in listener perception (Scott, Isard, & de Boysson-Bardies, 1985). Many definitions of rhythm across different fields still imply temporal periodicity, although periodicity is but one special case of rhythm. While periodicity is always rhythmic, rhythm does not always result in a periodic signal (Patel, 2008: 98). In contrast, broad definitions understand auditory rhythm as a systematic patterning of sound in terms of prominences (or accents, beats), their temporal distribution and grouping (Patel, 2008). Our previous study (Rathcke & Smith, 2015) put this definition to an empirical test and provided some evidence that linguistic rhythm and the surrounding controversy may indeed be related to the prosodic structure (i.e., phrasal prominences and boundaries), at least in languages that have lexical stress, such as English (Cutler, 2015; Hayes, 1989).

The present study adopts the view that rhythm in speech and language arises as a consequence of a general internal timekeeping mechanism that allows us to synchronize, anticipate, and adapt our behavior in response to an external stimulus (Repp, 2005). Similar to music (Patel, 2008), we expect speech and language rhythm to arise from temporal alternations between strong and weak elements as well as their

E-mail address: tamara.rathcke@uni-konstanz.de (T. Rathcke)

groupings into larger units (Rathcke & Smith, 2015). To study these properties, we deploy sensorimotor synchronisation – a methodological paradigm that examines rhythmic entrainment by means of perception–action coupling (Repp, 2005; Repp & Su, 2013). Auditory and motor systems interact closely when encoding rhythmic patterns of sound stimuli (Chen et al., 2008; Grahn & Rowe, 2009; Iversen et al., 2009; Morillon et al., 2014; Stupacher et al., 2013; Su & Pöppel, 2012; Zatorre et al., 2007). Behavioral and neuroimaging evidence has further shown that the connection between motor and auditory systems is bidirectional. On the one hand, movement seems to shape auditory rhythm perception in adult life and infancy (Phillips-Silver & Trainor, 2005, 2007). On the other hand, motor areas in the brain are activated when participants are passively listening to music rhythms (Chen et al., 2008; for a review, see Ross, Iversen, & Balasubramaniam, 2016). It has been suggested that moving in time with a sound engages the auditory-motor feedback loop which helps to find and track the beat (Su & Pöppel, 2012).

To date, the tight link between movement and rhythm perception has only infrequently been exploited in research on rhythm in speech and language (Rathcke, Lin, et al., 2021). Given that linguistic rhythm has been a matter of debates for a prolonged period of time (e.g., Kohler, 2009; Nolan & Jeon, 2014; White, Mattys, & Wiget, 2012), the current understanding of the rhythm structure in speech and language may benefit from insights of a movement-based paradigm (Rathcke, Lin, et al., 2021). The present study aims to shed light on this issue by testing what units of linguistic or acoustic structure entrain rhythmic movement with language.

1.2. Sensorimotor synchronization paradigms

Sensorimotor synchronization paradigms have been used to understand rhythm perception in external sounds for several decades (for a review, see Repp, 2005; Repp & Su, 2013). Sensorimotor synchronization refers to a type of motor behavior that is coordinated with an external, temporally structured event (Repp, 2005). In such paradigms, a sound (usually a metronome sequence or a music excerpt) is presented to the participants who are instructed to synchronize with the beat of the sound. The beat is defined as a pulse – that is, a regularly occurring, subjectively prominent event in time, leading to the perception of “an endogenous periodicity” (Large and Snyder 2009: 46). The presence of a perceivable beat is a prerequisite for sensorimotor synchronization which occurs even in complex, syncopated auditory prompts (Drake & Palmer, 1993; Gerhard, 2003; Large & Palmer, 2002; Lindbolm & Sundberg, 2007; London, 2004; Madison, 2014). Strong and weak pulses can alternate and thus lay the foundation for the metrical structure and rhythmic hierarchy (Large & Snyder, 2009).

Among existing sensorimotor paradigms, finger-tapping is one of the most widespread forms of eliciting synchronized movement (Repp, 2005). Participants’ sensorimotor behaviors in a finger-tapping task can be described using a number of metrics which include the variability of the inter-tap intervals and asynchronies between taps and acoustic onsets of rhythmic targets in the auditory stimulus. In general, synchronizing with an external beat is not a difficult task for participants without a clinical history of movement or other disorders, though musicians show improved performance in terms of smaller asynchronies (Aschersleben, 2002) and more stable tapping intervals (Repp, 2010), as compared to non-musicians (see Dalla Bella et al. (2017) for a review of individual tapping abilities).

Despite its popularity in rhythm perception research, synchronization paradigms have rarely been applied to the study of rhythm in language (Rathcke, Lin, et al., 2021; but see Wagner et al. (2019) for an alternative approach). One of the earliest studies to use a somewhat similar paradigm was conducted by Allen (1972). Three participants of his experiment were asked to tap on one predefined target syllable while listening to a sentence loop. The results showed that the collected taps fell close to the vowel onset of the target syllable. However, Allen’s pioneering research did not inspire many follow-up studies. Instead, the typology of linguistic rhythm classes (Abercrombie, 1967) and the (problematic) approach of rhythm metrics (Ramus, 2002; Ramus et al., 1999) have been dominating the field (for a critique, see e.g., Arvaniti, 2012; Rathcke & Smith, 2015).

Very few recent studies have piloted and elaborated the sensorimotor synchronization paradigms with speech. Lidji et al. (2011) recruited monolingual English and French speakers as well as English-French bilinguals and asked them to tap along with English and French sentences which contained monosyllabic words and had a predominantly trochaic, strong–weak stress pattern at the phrasal level (e.g. SW SW SW S as in ***Days*** *are* ***bright*** *with* ***lots*** *of* ***light***). The participants of the study listened to three repetitions of each sentence and were instructed to tap along with the second and the third repetition, without being given predefined synchronization targets (unlike in Allen’s (1972) design). The results showed that all participants, regardless of their native language, tapped more regularly to English than to French sentences. In addition, English monolinguals were more likely to tap to every two or four words, while French monolinguals tended to tap to every word in both languages. Continuing this line of research, Leong and Goswami (2014) tested the synchronization ability of dyslexic and control participants with English nursery rhymes. The rhymes were also repeated three times, and participants were asked to tap along with each repetition and maintain their tapping patterns during interspersing silence intervals, though only the taps coinciding with the second and the third repetition of each rhyme were included in the analyses. Both dyslexic and control participants of the study tapped along with every stressed (but not unstressed) syllable of the rhymes, without being instructed to do so. Falk, Rathcke, and Dalla Bella (2014) used a finger-tapping synchronization task with sentences whose timing properties were artificially altered and found that participants’ tapping patterns changed in response to the subtle timing variability. Finally, our most recent work further showed that participants tended to synchronize with vowel onsets when asked to tap along with the beat of a sentence played on a 20-repetition loop, without being given any explicit instruction about the temporal targets of the taps (Rathcke, Lin, et al., 2021).

Just like people can naturally synchronize with the beat of music (or with the beat of a metronome in the lab), the above-mentioned studies indicate that synchronization with speech is possible if spoken sentences are repeated, though the variability of tapping intervals produced with speech is admittedly larger than it is with a metronome or music (Dalla Bella et al., 2013). However, the origin of such increased variability in synchronization behavior can be two-fold: (1) it can result from the difficulty to discern a clear beat in speech, or (2) it can stem from an inherent temporal variability of natural speech (see Rathcke, Lin, et al. (2021) for a discussion). Moreover, previous research has shown that vowels are the most likely anchors of sensorimotor synchronization in both natural speech (Rathcke, Lin, et al., 2021) and artificial verbal prompts (Rathcke et al., 2019). Vowels frequently constitute syllabic nuclei and are the only obligatory element of a syllable in many languages of the world (Féry, 2003; Fudge, 1969; Levin, 1985; Pike & Pike, 1947). However, in some languages – among them English – consonants can also form syllabic nuclei, even though this option is often restricted to unstressed syllables only (Pouplier & Beňuš, 2011; Ridouane, 2002; Toft, 2002). Thus, it remains unclear if the target of sensorimotor synchronization in speech is specifically the vowel, or more generally, the syllable nucleus.

Empirical evidence to clarify this issue would help to answer a broad theoretical question if rhythmic attention in speech is guided by the fluctuations in vocalic energy alone (cf. Fowler & Tassinary, 1981; Marcus, 1981; Morton, Marcus, & Frankish, 1976), or alternatively, locks on the linguistic structure of the prosodic hierarchy that consists of interconnected layers of increasingly larger constituents, with the syllable representing the lowest level of the hierarchy (Hayes, 1989; Langus et al., 2012; Selkirk, 1984).

The present study seeks to provide new evidence on these questions, by investigating rhythmic synchronization with natural speech, and by examining acoustic landmarks that serve as reliable anchors of stable synchronization. Following previous synchronization paradigms that were developed for musical and other rhythmic stimuli (e.g., Sowiński & Dalla Bella, 2013), we measure signed and absolute asynchronies between finger taps and several linguistic or acoustic landmarks. The magnitude of temporal proximity between taps and landmarks can provide insights into rhythm perception in speech. The closer a tap is to a specific acoustic landmark, the more likely that landmark is perceived as the beat anchor. We further estimate tapping probabilities that indicate how likely a tap is to be produced near an acoustic landmark (Rathcke, Lin, et al., 2021; cf. Repp, 2004). The higher the tapping probability found for a specific acoustic landmark, the more likely the landmark is perceived as the beat anchor. Both measurements have been used in our previous work and provided preliminary evidence for the domain of beat perception in speech (Lin & Rathcke, 2020; Rathcke & Lin, 2021; Rathcke et al., 2019; Rathcke, Lin, et al., 2021). In addition, we also compare the variability of measured sensorimotor behaviors with the different landmarks. Following Leonard and Cummins (2011), we assume that stable synchronization with a specific landmark is indicated by a reduced variability of SMS measured with that landmark.

1.3. Potential anchors of sensorimotor synchronization in speech

1.3.1. Nucleus onsets

Previous research has shown that the vowel onset is likely to constitute the anchor of sensorimotor synchronization with speech when participants are synchronizing with a predefined syllable in natural speech (Allen 1972), with isolated monosyllabic words (Rathcke et al., 2019), or with naturally spoken sentences (Lin & Rathcke, 2020; Rathcke, Lin, et al., 2021). This research reliably established that vowel onsets outperform syllable onsets in predicting the location of participants' taps (Allen, 1972; Lin & Rathcke, 2020; Rathcke et al., 2019; Rathcke, Lin, et al., 2021).

However, previous research has not addressed what might constitute a synchronization anchor in syllables whose nucleus is not a vowel. In many languages, syllable nuclei can contain a sonorant (as in English 'but<u>ton</u>' [bʌtn] or in German 'ha<u>ben</u>' [ha:bm]), albeit only in unstressed syllables. Sonorants such as nasals and laterals are permissible in English syllable nuclei, though acoustically, they tend to have lower energy than vowels (Ladefoged, 1993). Comparing SMS with the two types of syllable nuclei (vocalic vs. non-vocalic) can therefore help uncover to what extent rhythmic attention in speech is guided by the structural unit of a syllable or rather by the properties of amplitude envelopes and energy changes related to vowel onsets (e.g., Goswami & Leong, 2013; Prasanna, Reddy, & Krishnamoorthy, 2009). If acoustic properties of vowels lay foundation to rhythmic beat tracking in language, sensorimotor synchronization with non-vocalic nuclei is expected to be limited or non-existent. If rhythmic attention is primarily linguistic, guided by the top-down knowledge of the linguistic structure, we expect no difference between sensorimotor synchronization patterns with unstressed syllables containing either vowels or sonorant consonants. And finally, a combined influence of the linguistic knowledge and the acoustic cues is also likely. In this case, we expect subtle differences between the two types of nuclei with regards to tapping probabilities, stability and/or asynchrony.

1.3.2. Perceptual centers

The notion of the 'perceptual center' or the 'P-center' has been put forward to account for the repeated finding that acoustic syllable onsets do not seem to coincide with perceived syllable onsets (Marcus, 1981; Morton et al., 1976; Scott, 1998). Accordingly, temporal regularities in natural speech are difficult to establish because they follow P-centers and not acoustics of realized linguistic units such as syllables. Several studies have proposed that the P-center lies close to the vowel onset (Allen, 1972; Šturm & Volín, 2016) or suggested that it is located somewhere between the onset of the syllable voicing and the local energy spike (Cummins & Port, 1998). However, the exact location of the P-center has remained difficult to establish (Villing et al., 2011). It seems to be influenced by numerous factors, including presence and number of the preceding and the following consonants (Allen, 1972; Marcus, 1981; Morton et al., 1976; Pompino-Marschall, 1989), risetime of the amplitude envelope (Scott, 1998), and the energy distribution of the syllabic amplitude envelope (Howell, 1988). One of the most explicit, formalized approaches to the identification of P-centers developed to date is the algorithm by Šturm & Volín (2016). It considers properties of (smoothed) energy contours and measures their local maxima (*maxE*) and moments of the fastest energy increase in the proximity of a vowel (*maxD*). The algorithm showed promising results in a study with Czech participants whose task was to synchronize production of bisyllabic word sequences with a metronome. The algorithm-generated *maxD*-measure of participants' speech productions was in close alignment with the metronome beats, which was then interpreted as evidence for an adequate P-center location.

Given that the issue of P-centers has mostly been studied in non-natural settings such as isolated syllables, words, or isochronous speech (Hawkins, 2014), the general applicability of the idea to more natural speech materials requires further investigation. Our pilot studies compared sensorimotor synchronization with *maxD* and *maxE* landmarks that can be derived by the algorithm (Lin & Rathcke, 2020; Rathcke, Lin, et al., 2021). In line with previous findings (Šturm & Volín, 2016), *maxD* was a clear winner. Specifically, the results showed that both *maxD* and vowel onsets were equally successful at predicting the location of sensorimotor synchronization peaks in natural speech (Lin & Rathcke, 2020; Rathcke, Lin, et al., 2021), though the materials of our pilot studies were limited in number and contained only vocalic, but not consonantal, syllable nuclei. Considering the comparable predictive power of the two best-performing landmarks, automatically derived *maxD* might be preferrable over manually annotated vowel onsets in the context of a larger dataset since *maxD* requires less manual labor and may be less affected by occasional segmentation difficulties (Turk et al., 2012). The algorithm operates within syllable boundaries and has the potential to be fully automatic, if combined with a suitable syllabification annotation algorithm (e.g., Marchand et al., 2009).

1.3.3. Further acoustic landmarks

Since vowel onsets can be difficult to segment in certain contexts, most notably after sonorant consonants (Turk et al., 2012), and since *maxD*-algorithm relies on manually defined syllable boundaries to run, the need for a fully automated approach to the identification of acoustic landmarks is relatively high. In addition to the landmarks discussed above, there might be further automatically derivable landmarks to optimally represent the anchor of sensorimotor synchronization and the domain of rhythmic attention in speech. A list of such potential landmarks is discussed below.

1.3.3.1. Pitch and periodic power. Some accounts of speech rhythm highlight the fact that pitch is an important cue for prominence and should thus be one of the key factors shaping rhythm perception in speech (Arvaniti, 2009; Kohler, 2008). This idea has not been extensively tested in previous research, though there is some indication that the perception of rhythmic variability can be shaped by pitch dynamics (Dilley & McAuley, 2008; Morrill et al., 2014; Niebuhr, 2009). In a pilot study, we examined sensorimotor synchronization with high and low tones on different timescales (Rathcke et al., 2019). High tones (260 Hz) and low tones (130 Hz) were presented either fast (at 300 ms inter-onset interval, IOI), or slow (at 600 ms or 1200 ms IOI). Participants were instructed to tap along with either high or low tones in each sound sequence. Our results showed that the synchronization with low tones was more accurate than the synchronization with high tones at the same timescale, indicating that pitch may influence beat perception as measured by sensorimotor synchronization. The results from a parallel line of research are, however, mixed. Researchers who study speech-accompanying gestures argue that the gesture apices tend to coincide with F0-peaks of locally prominent events in speech, while others do not agree (for a review, see Esteve-Gibert & Prieto, 2013). Neuroimaging research has suggested that there are some distinct but also some common brain regions that are activated during processing of pitch vs. timing properties in spoken nonce stimuli (Zhang et al., 2010).

The idea of combining pitch and periodic power, the Pitch and Periodic Power algorithm (PPP), originates in the work by Albert et al. (2018). It has been successfully utilized for creating an empirically useful way to visualize perceptually salient pitch sections in speech (Cangemi et al., 2019). While the algorithm neither categorizes prominence *per se* nor gives any threshold values that could be used to examine prominence alternations in speech, it highlights those parts of F0-tracks that will be perceptually more salient, given their periodic power distribution (a log-scaled derivative of acoustic intensity).

In a previous pilot study, we adopted the PPP algorithm and generated a series of landmarks whose location coincided with equally weighted, highest local F0-frequency and periodic power. However, the landmark performed rather poorly at predicting sensorimotor synchronization behaviors of native English participants who took part in the study (Lin & Rathcke, 2020). The question that remained open and that we now seek to clarify is if, and to what extent, sensorimotor synchronization with speech is guided by the properties of pitch contours. Accordingly, the present study aims (1) to systematically vary pitch-to-power ratios, modifying the relative contribution of pitch to the generation of acoustic landmarks within a syllable, and (2) to compare the landmarks' potential to predict synchronization peaks. In addition to the algorithm-derived PPP ratios, the moderating role of the phonological structure of tunes will be investigated (Ladd, 2008). That is, we will examine if the presence of phonological pitch events such as pitch accents and (initial/final) boundary tones may improve the capture of sensorimotor synchronization with PPP-derived, acoustic landmarks. Given that speech-gesture synchronization in natural conversational speech is influenced by the presence of F0-peaks that express pitch accents, we hypothesize that taps will be produced closer to the PPP-derived landmarks in those syllables that coincide with the phonologically relevant pitch events.

1.3.3.2. Multiple timescales. Some proposals suggest that speech rhythm has a high level of complexity which cannot be adequately modeled by incorporating just one dimension of timing, since each rhythmic level is likely to interact with others (Arvaniti, 2012; Cummins, 2009; Kohler, 2008; O'Dell & Nieminen, 1999; Tilsen, 2009). To this end, Tilsen and Arvaniti (2013) outline a possible approach of modelling rhythmic complexity at multiple timescales – the empirical mode decomposition (EMD).

EMD is a data-driven method to acquire the information about inherent periodicities hidden in non-linear and non-stationary data (Huang et al., 1998). A signal is decomposed through a sifting procedure by subtracting the mean of the upper and lower amplitude envelope from the original signal until the mean is close to zero. This procedure acquires an intrinsic mode function (IMF) which is then subtracted from the original signal. Starting from the newly acquired signal, the sifting procedure repeats again to find the next IMF. The first IMF is assumed to capture periodicity on the shortest timescale and every subsequent IMF moves up to periodicities at a subsequently larger timescale. Using EMD, Tilsen and Arvaniti (2013) consider multiple periodicities in the acoustic signals of connected speech and suggest that the syllabic and supra-syllabic (or inter-stress) periodicities can be represented by the first and the second IMF, respectively. They piloted the application of the EMD method on amplitude envelopes of speech samples taken from a corpus of English speech and then applied the method to a cross-linguistic corpus of read and spontaneous speech containing samples of English, German, Greek, Italian, Korean, and Spanish recordings. The results showed that the power distribution, rate, and rhythmic stability metrics derived from the first and the second IMF can differentiate languages and speech styles (read vs. spontaneous).

Following on from this line of research, our pilot study examined the ability of IMF-maxima to predict participants' synchronization (Lin & Rathcke, 2020). However, the results showed that both IMF-1 and IMF-2 maxima performed worse than other landmarks derived for a small set of English sentences (Lin & Rathcke, 2020). Despite this result, IMF remains appealing as one of the automated procedures that requires the least amount of supervision. To further examine the potential of the EMD-method in the context of sensorimotor synchronization, the present study includes both IMF-maxima and IMF-minima as potential anchors of rhythmic movement with speech.

In summary, our previous work on the acoustic landmarks that might attract and anchor rhythmic movement with natural speech (Lin & Rathcke, 2020; Rathcke, Lin, et al., 2021) can be further extended to include additional, improved landmark locations and also a larger set of speech materials with varied syllabic nuclei.

1.4. Aims of the present study

Research reviewed in the introduction has demonstrated that sensorimotor synchronization, and specifically the finger-tapping paradigm, provides a valuable tool for studying the nature of rhythmic attention in speech and language. However, a comprehensive evaluation of the acoustic landmarks that may serve as synchronization anchors and thus predict tap locations in natural, diverse linguistic materials is still missing. Moreover, the minimal unit of rhythm has not been fully established as previous research concentrated exclusively on simple linguistic structures comprising syllables with vocalic nuclei only. The present study addressed the issue by running an experiment with the following aims:

(1) To examine the overall contribution of pitch to creating stable acoustic anchors of sensorimotor synchronization in natural speech and to establish the best pitch-to-power ratio that can best predict tap locations. Eleven ratios from pitch only (PPP 0:10) to power only (PPP 10:0) are examined (Fig. 1).
(2) To test both the onsets of sentence-internal periodicities (represented by local minima of IMF-contours) and the midpoints of the corresponding timescales (represented by IMF-maxima) in their potential to predict tap locations in natural speech, for both syllabic (IMF-1) and inter-stress (IMF-2) timescales.
(3) To compare synchronization patterns with manually identified nucleus onsets vs. with algorithm-generated acoustic landmarks. In extension of our previous research, the special focus of the present study lies on the comparison between syllables containing vocalic vs. consonantal nuclei, establishing if vowels (as opposed to syllable nuclei) constitute the primary target of rhythmic attention in natural speech.

2. Method

2.1. Stimuli

Twenty English sentences were recorded at sampling rate of 44,100 Hz in a sound-attenuated booth of the linguistics laboratory at the University of Kent. The sentences were produced by a native 30-year-old female speaker of Standard British English (see the full sentence list in Table 1 and the example sound files on OSF, https://osf.io/7gejt/). Sentences varied in the number of syllables (ranging from 7 to 13) and duration (ranging from 1.2 to 2.3 seconds). Each sentence was repeated 10 times in the main experiment and 20 times in the practice session, with a 400 ms pause between repetitions. The number of repetitions was determined by the results of our previous experiments in which we observed that participants' tapping asynchronies were mostly stable after five repetitions (Rathcke, Lin, et al., 2021).

2.2. Identification of acoustic landmarks

2.2.1. Manual annotations

An expert phonetician (the first author) manually segmented and labelled syllable and nucleus boundaries. Nucleus onsets were demarcated by the onset of voicing and a continuous formant structure in post-obstruent contexts (i.e., vocalic intervals excluded pre- and post-aspiration as well as glottalization). The acoustic and auditory criteria were combined in post-sonorant contexts. A similar procedure was applied to the annotation of consonantal nuclei with the added consideration of reduced intensity (for annotation examples see Praat TextGrid files on OSF, https://osf.io/7gejt/). Syllable boundaries were identified following the maximal onset principle (Kahn, 2015). The materials (see Table 1) posed no syllabification issues. By and large, the distribution of unstressed syllables with consonantal vs. vocalic nuclei was balanced across prosodic conditions as they occurred either in word/phrase-medial or word/phrase-final positions, though the total count of consonantal nuclei was lower than the count of vocalic nuclei in these (naturalistic) sentences.

Manual annotations further included the tune structure of the test sentences. The first author (trained in perceptual annotation of phonologically relevant pitch events) followed general guidelines for the annotation of the Tone and Break Indices (ToBI, Silverman et al., 1992) and added tonal specifications to the relevant syllables. The annotation combined perceptual criteria of determining phrasal prominence and breaks with a visual examination of acoustic F0-tracks. Examples of the annotations can be found on OSF (https://osf.io/7gejt/). A great diversity of pitch accents and boundary tones was produced by the speaker, including L*, L*+H,!H + L*, H + L*,!H*,^H*, H*, L + H*, %L, L-L%, %H, L-, H- (see Table 2). For the purposes of statistical modelling, we reduced the annotations to key pro-

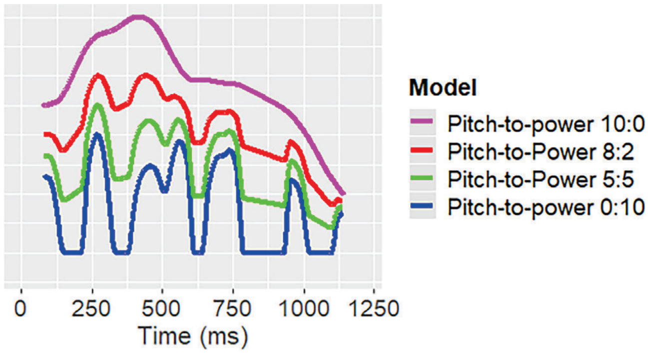

Fig. 1. An example of different pitch-to-power ratios for the test sentence "*Her cackle was not subtle*". The top pink line is the interpolated F0, while the bottom blue line is periodic power. The red and the green lines in-between are the combination of pitch and periodic power in different ratios (i.e., it represents the sum of the normalized values of pitch and periodic power that is multiplied by the given ratio). Please note that pitch and amplitude of the models (plotted along the schematic Y-axis) have been adjusted to avoid overlapping.

Table 1
The list of sentences used in the present study. The consonantal nuclei are marked in bold. Underlined syllables were expected to contain a sonorant consonant but were produced with a schwa, instead.

Sentence	Total number of syllables	Number of non-vocalic nuclei
Her cac**kle** was not sub**tle**.	7	2
The ca**ttle** sudd**en**ly escaped.	8	2
The broken bo**ttle** made a ra**ttle** as it fell.	12	2
The milk carton is biodegrada**ble.**	11	1
The coup**le** watched the stars twink**le**.	8	2
Tom and his un**cle** explored the jung**le**.	10	2
Fresh app**le** juice is delicious.	8	1
The seamstress seldom lost her butt**ons**.	9	1
The only juice available was pineap**ple**.	12	1
Passers-by had tro**dden** down the autumnal leaves.	12	1
The café will cer**tain**ly serve bacon on toast.	12	1
The stu**dent** had wri**tten** an exceptional essay.	13	2
He felt su**dden** pain in his an**kle**.	9	2
The woo**den** chair was not afforda**ble**.	10	2
Grandma's rhubarb crum**ble** is heaven.	10	1
In the autumn, she wears comforta**ble** boots.	11	1
There are lots of pu**ddles** in autumn.	9	1
The pur**ple** bicy**cle** was the fastest.	10	2
The door han**dle** has broken.	7	1
The peo**ple** were woken by thunder.	9	1

sodic phenomena. First, we distinguished between units that were unaccented (133 syllables) and accented (63 syllables), edge-delimiting (50 syllables) and phrase-medial (146 syllables). Second, each tonally specified syllable was classified as belonging to the group of either high or low tones and subdivided according to its function into prominence-lending or edge-delimiting (see Silverman et al., 1992, for more details). Given the basic nature of the tonal specifications used here (tone class only, see Table 2), a cross-annotator check was deemed unnecessary.

Table 2
Summary of tonal annotations of the study materials following ToBI-guidelines (Silverman et al., 1992).

Tonal function	Tone class	N syllables	Tonal specifications
Prominence-lending	High	35	H*, L + H*,!H*,^H*,
	Low	28	L*, L*+H,!H + L*, H + L*
Edge-delimiting	High	8	%H, H-
	Low	42	%L, L-, L-L%

2.2.2. Maximal energy increase – maxD

Procedures for calculating the local timepoint of the maximal energy increase (*maxD*) were adapted from previous work (Šturm & Volín, 2016). First, a raw energy contour of each sentence was derived by squaring raw amplitudes and using a 40-ms window and a 1-ms shift. A smoothed energy contour was then acquired by applying a 6th-order moving average filter to the raw energy contour. In a subsequent step, a raw energy-difference contour was created by calculating the difference between two adjacent samples of the smoothed energy contour ($E_T - E_{T\ 1}$) and further smoothing via a 10th-order moving average filter. Examples of the two smoothed contours derived from one of the sentences are shown in Fig. 2. For each syllable, *maxD* was identified as the highest local value in the smoothed energy-difference function (indicated by arrows in Fig. 2).

2.2.3. Pitch and periodic power – PPP

The approach developed by Albert, Cangemi, & Grice (2018) was adapted for the purposes of the present study and used to calculate different ratios between periodic power and pitch. The F0 and intensity analyses were performed in Praat (Boersma & Weenink, 2018). Periodic power reflects intensity values at each given timepoint of a sentence that has measurable F0 within a predefined range of values between 40 and 600 Hz. The settings for the extraction of intensity values (measured in dB) were at default except for the pitch floor that was decreased to 40 Hz. The *"subtract mean pressure"* function, and the default time step (0.8/minimum pitch) were applied. The F0 analysis was conducted by an autocorrelation method on the Gaussian-window signals using the default F0-pathfinder setting in Praat (Boersma & Weenink, 2018). The silence threshold was set at 0.03, the voicing threshold at 0.45, the octave cost at 0.01, the octave jump cost at 0.35, and the voiced/unvoiced cost at 0.14. The resulting F0-contour was subsequently smoothed by a 10-Hz low-pass filter. For the calculation of periodic power, the intensity value was first converted from the sound pressure level measured in dB into the amplitude contour (see (1) for the equation), then the intensity value was multiplied by a weighting quotient (from 0 to 1) that described the associated F0-value acquired from Praat.

$$4 \times 10^{\ 10} \times 10^{(sound\quad pressure\quad level/10)} \tag{1}$$

At each timepoint (here, the information was sampled every millisecond), the calculated periodic power reflects local intensity with reference to the presence of concomitant F0. Accordingly, F0 values of periodic power omit voiceless segments (see Albert et al., 2018). The final output of periodic power contour was logarithmically transformed (set to 0 if negative) and smoothed using a local regression method with a span of 0.03.

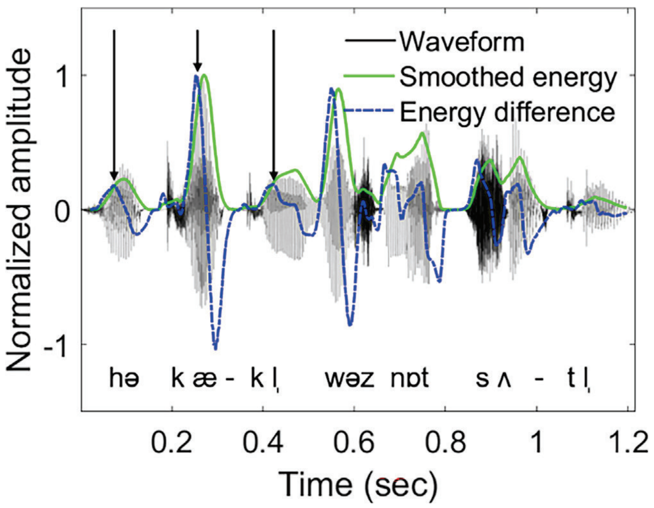

Fig. 2. Waveform (dark grey lines), smoothed energy contour (solid green lines), and smoothed energy difference contour (dashed blue lines) for the sentence "*Her cackle was not subtle*". The timepoints of the *maxD*-landmarks are indicated by black arrows (the first three syllables only).

In order to define possible acoustic landmarks for the purposes of this study, we created different representations of the weighted relationship between F0 and periodic power. Note that F0 and periodic power were both normalized with reference to the maximum F0 and the maximum power of the corresponding sentence. Accordingly, the sum of F0 and periodic power was represented by different ratios of the two components. This was done by their summation using a 10%-step increment, starting from 0% F0 and 100% power (0:10-ratio) and going up to 100% F0 and 0% power (10:0-ratio), which resulted in a total of eleven pitch-to-power ratios (PPP; see Fig. 1 for an example of different PPPs). For the synchronization analyses, we calculated timepoints of syllabic PPP-maxima for each of the PPP contours (see Fig. 1 for an example of different PPPs).

2.2.4. Intrinsic mode function - IMF

Acoustic landmarks associated with different timescales of the signal periodicity were derived using the EMD approach as discussed by Tilsen and Arvaniti (2013). The landmarks were based on the IMFs (Huang et al., 1998). Accordingly, the raw sound signal was first passed through a 4th-order Butterworth band-pass filter of 400–4000 Hz, to minimize both the contribution of F0 (lower cut-off) and that of sibilants (upper cut-off). The resulting signal was then subjected to a 4th-order Butterworth low-pass filtering at 10 Hz with an assumption that the syllabic periodicity fluctuates on a scale of 100 ms or more. This type of filtering highlights the properties of sonorous energy in the original signal, which is somewhat similar to periodic power derived via the PPP-procedure.

The main interest of the present study lies in the precise temporal location of periodic timescales while previous work that applied IMFs to the study of linguistic rhythm focused primarily on the periodic frequencies arising at the identified timescales (Tilsen & Arvaniti, 2013). To improve the localization of the timescales identified as IMF-1 and IMF-2, we applied an additional realignment step to the filtered signal (not performed in our previous work, Lin & Rathcke 2020). The overall temporal delay of the filtered signal was temporally adjusted back to align with the smoothed energy contour (Fig. 2). That is, the first peak of the low-pass filtered signal was realigned to the first peak of the energy contour. An average forward-shift for the 20 sentences was 44.2 ms (*SD* = 7.7 ms). This was close to the approximate estimation of the delay time at zero frequency (41.6 ms) for a 4th-order Butterworth low-pass filter (Manal & Rose, 2007). The realigned signal was then normalized by subtracting the mean amplitude of the signal from the local amplitude at each timepoint, with a rescaling from −1 to 1. Finally, the signal was downsampled by factor 100 with the application of a Tukey window (*r* = 0.2, for more detail and rationale see Tilsen & Arvaniti, 2013).

After the signal was pre-processed through the above procedures, the EMD was applied, decomposing the signal into its component IMFs through a sifting procedure (Huang et al., 1998). During the sifting, an amplitude envelope was first acquired by connecting local minima and maxima through a cubic spline interpolation. The mean of the upper and lower amplitude envelope was then repeatedly subtracted from the original envelope until its mean was close to zero. The algorithm checked that the number of zero crossing and amplitude extrema at each given timepoint were either equal or differed no more than by factor 1.0. Once these requirements were met, an IMF was identified. Subsequently, the IMF was subtracted from the original signal and the sifting procedure was repeated once more to find the next IMF (Tilsen & Arvaniti, 2013).

The output of the EMD-analysis contained a set of IMFs which represented periodic oscillations of energy in the signal, with the periodicity of the IMF timescales being determined by the properties of the signal itself. Previous research has indicated that the first and the second IMF may represent the syllabic and interstress interval fluctuations of natural speech, respectively (Tilsen & Arvaniti, 2013). The local onsets (IMF-minima) and midpoints (IMF-maxima) of the two timescales were obtained from IMF-1 and IMF-2. The selection of the timepoints was guided by two considerations: (1) each IMF-minimum or maximum had to be at least 20% of the highest maximum (or lowest minimum) in the corresponding sentence, and (2) adjacent maxima (or minima) had to be at least 100 ms apart.

2.2.5. Summary and overview

The approach outlined above yielded seven acoustic landmarks for our study of sensorimotor synchronization with natural speech. Manually identified nucleus onsets and algorithm-generated *maxD* represented two approximations of the P-center location discussed in the literature (Šturm & Volín, 2016) while the timepoints identified via IMF and PPP procedures signified potentially relevant properties of the energy contours with a greater or a lesser contribution from the local F0 information.

Fig. 3 displays an example of the seven landmarks generated for the word *juice* (taken from the test sentence "*Fresh apple juice is delicious*"). As can be seen, the landmark locations varied a lot. Some of them were closer to the nucleus onset while others were not.

2.3. Task, procedure, and apparatus

Prior to the experiment, participants did a practice session with six trials and had a chance to clarify their questions with the experimenter. They were instructed to "feel the beat" of looped sentences and to start synchronizing with the beat as soon as they felt ready, by tapping on a drum pad with the index finger of their dominant hand. The stimuli of the practice session contained 20 repetitions of each sentence in a fixed order and were not used in the main experiment that followed. In the main experiment, participants had to listen to the first two repetitions of each sentence, and then to start synchronizing with the sentence beat from the third repetition onwards until the looped presentation stopped. The presentation of the next sentence loop began after a longer pause. All sentences of the main experiment were repeated 10 times and played in a randomized order. The experiment was conducted in a quiet testing booth of the linguistics laboratory at the University of Kent. One experimental session took approximately 30 minutes to complete.

Synchronization data were collected via a drum pad (HPD-20, Roland) and a Dell laptop running Cakewalk software (Cakewalk by Bandlab, 2019). The laptop controlled the sound playback and recorded participants' taps in MIDI format. The sound was presented to the participants via good-quality headphones (Sennheiser HD 380 Pro).

2.4. Participants

Thirty-two monolingual, native speakers of British English participated in the experiment (21 females; *mean age* = 20.4 years, *SD* = 2.5 years). None of them self-reported any language- or movement-related disorders. All participants gave written consent and received a small remuneration for their time. The study was approved by the ethics committee of the University of Kent.

2.5. Tapping data pre-processing

The collected tapping data were extracted using the MIDI toolbox in MATLAB (Eerola & Toiviainen, 2004). First, a 5-ms recording latency of the equipment was subtracted from the raw timepoints of tap locations.[a] Next, all taps collected for each sentence loop were split up into the corresponding repetition cycles. Taps collected up to 300 ms prior to the repetition onset were assigned to the upcoming repetition cycle, given the time window of prediction of upcoming rhythmic events measured in previous entrainment studies (Repp, 2005; cf. Rathcke, Lin, et al. 2021). Any taps produced before the third repetition cycle and/or after the 10th repetition cycle were excluded from further analyses.

A tapping distribution was then generated for each sentence and participant, by aggregating the taps produced during eight repetitions (from the third until the tenth loop cycle) and by applying a Gaussian kernel density estimation with a 1/8 bandwidth in R. The procedure of probability density estimation involves the application of kernel-based smoothing which answers to the problem of statistical inference about population-level distributions based on a finite sample of data points (e.g., Kolassa, 2020). It has been used in related SMS-studies that investigated synchronization of body movement with auditory signals of a high complexity level like contemporary music of different genres (Dotov et al., 2021). Peaks in the tapping distributions were identified as the local density maxima. To count as a local maximum, a density peak had to reach at least 40 percent of the largest density maximum identified for the given distribution (see Fig. 4). A window of 100-ms between adjacent tapping peaks was employed, to account for physiological limitations in synchronized movement (Repp, 2005). If two tapping peaks were identified within a given 100-ms interval, the highest of the two was chosen.

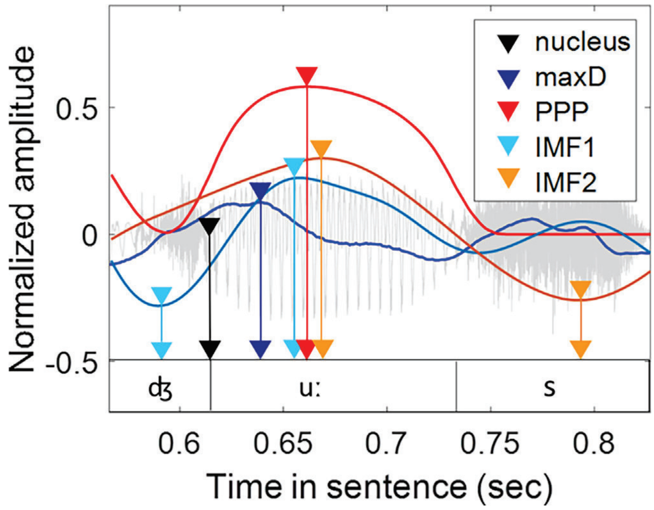

Fig. 3. Waveform (grey lines), segmentation, and a broad phonetic transcription of the word *juice*. Colored arrows indicate the locations of the seven landmarks investigated in the present study: nucleus onset (black triangles), *maxD* (dark blue triangles), PPP (red triangles, exemplified by a 5:5-ratio here), maximum and minimum of IMF-1 (light blue triangles), maximum and minimum of IMF-2 (orange triangles).

2.6. Calculation of tapping probability, asynchrony, and variability

The timepoint of each acoustic landmark was tested as a potential anchor of the closest tapping peak for each participant and sentence (see Fig. 4). Following previous work on physiological limits (Repp, 2006) and temporal integration (Repp, 2004) in sensorimotor synchronization that was corroborated in our pilot studies with natural speech (Lin and Rathcke, 2020; Rathcke et al. 2021), a tapping peak found within a ±120 ms window around an acoustic landmark was treated as anchored to that landmark. If there was no tapping peak within the predefined time window, a non-anchoring response (0) was recorded for the corresponding landmark. Otherwise, the tapping probability was recorded as 1. Note that any given tapping peak could appear as being anchored to several neighboring landmarks (and have the probability of 1) as long the temporal requirement of the ±120 ms window was satisfied. This procedure allowed us to determine the potential of different landmarks to capture the tapping distribution.

Subsequently, asynchronies between tapping peaks and the nearby landmarks were calculated as a measure of synchronization error. Absolute asynchrony, i.e., the mode of the

[a] The latency was measured by comparing the time lag of sounds recorded simultaneously by CakeWalk and by a portable Zoom H4N recorder.

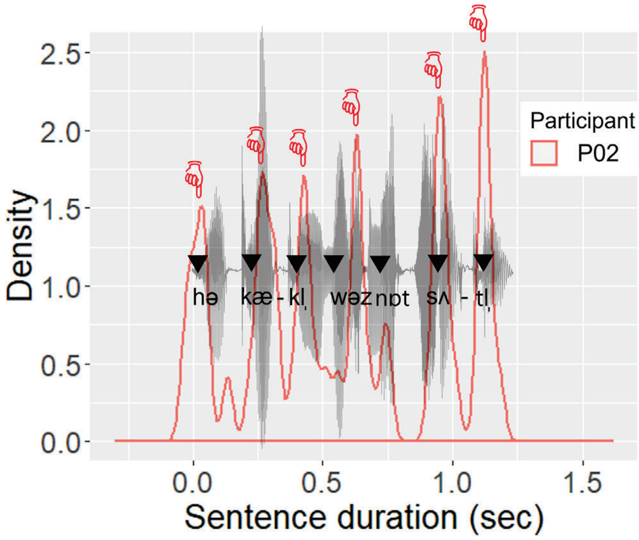

Fig. 4. Waveform (black lines) and nucleus onsets (black triangles, showing vowel onsets as an example of a landmark) of the test sentence "Her cackle was not subtle". Tapping density function obtained for participant P02 is plotted in orange as an example. Locations of tapping peaks are indicated by a tapping hand symbol. The absolute asynchrony is the mode of the interval between each tapping peak and its adjacent landmark (Lin and Rathcke, 2020; Rathcke et al., 2021, see text for more detail). Light blue circles indicate density maxima that did not reach the threshold of 40% of the highest peak and were therefore excluded from the asynchrony calculation.

interval between a tapping peak and a landmark, was used in this study to assess the movement-anchoring ability of each landmark. Tapping variability was measured as standard error of signed asynchronies calculated for each participant, landmark, and sentence independently.

2.7. Statistical analyses

Several linear and logistic mixed-effects regressions were performed. Three dependent variables were tested, *asynchrony, variability,* and *probability.* Asynchronies were measured in milliseconds and logarithmically transformed for statistical analyses, to deal with the typical distributional skewness of durational data (Baayen, 2008: 31ff). Variability was measured as standard error of signed asynchronies (in milliseconds) and also log-transformed. Probability described the likelihood of a tapping peak being detected (1) or absent (0) in the proximity of a landmark. *Participant* (1–32), *sentence* (1–20), and serial *order* of the landmark in the sentence (1–12) were fitted as the random intercepts (no models with random slopes converged). Three sets of predictors were tested on these data.

i. One set of models tested different ratios of F0 and periodic power. Eleven different pitch-to-power ratios were tested, from 0:10 (periodic power only) to 10:0 (F0 only) and fitted as a linear predictor (covariate) to the asynchrony data. The influence of tonal specifications on asynchronies with the PPP-landmarks was examined separately.
ii. Another set of models tested if tapping peaks were located consistently closer to the local IMF-maxima or the local IMF-minima at the syllable-level (IMF-1) or at the stress-level (IMF-2) periodicity.
iii. An additional set of models was performed to compare the anchoring ability of the algorithm-generated acoustic landmarks to the manually identified nucleus onsets. Best performing PPP and IMF landmarks from the analyses outlined in i. and ii. were used, in addition to *maxD* and manually annotated landmarks. Logistic mixed-effects models were performed to explain tapping probabilities in the proximity of the landmarks. To combat model convergence issues, the default optimizer of the logistic models was changed to *bobyqa* and the number of iterations increased to 100,000.
iv. A separate analysis addressed the issue of the best landmark that could predict synchronization with vocalic vs. consonantal nuclei by examining absolute asynchronies, tapping probabilities, and variability of sensorimotor synchronization. Permutation tests were additionally run to confirm the robustness of the results from linear mixed-effects models, given that the dataset containing consonantal nuclei was relatively small.

All analyses were performed in RStudio (running R version 4.1.2, R Core Team 2022).

3. Results

3.1. PPP ratios and the role of pitch

To examine the relative contribution of F0 to predicting the location of taps, the interaction between PPP with 11-factor levels and accents (accented vs. unaccented) was firstly fitted to the absolute asynchrony data. The best-fit model yielded a significant main effect of pitch-to-power ratios (β = 0.0008, *STE* = 0.0002, *t* = 3.18, *p* =.001), indicating a linear relationship between asynchronies and PPP while neither an interaction with accentuation (β < 0.0001, *STE* = 0.0003, *t* = 0.16, *p* =.88) and nor a main effect of accentuation (β = −0.03, *STE* = 0.18, *t* = −1.67, *p* =.10) were significant. These results are plotted in Fig. 5A (note that PPP of 10:0 eliminates periodic power from the landmark calculation, PPP of 0:10 eliminates pitch from the calculation while PPP of 5:5 gives pitch and power equal weighting). The linear relationship between asynchronies and PPP plotted in Fig. 5A holds regardless of the presence or absence of phrase-level accentuation and indicates that the lower the F0-ratio in the landmark location, the smaller the asynchronies. That is, the prediction of the tap location improved when the contribution of F0 to the calculation of the landmark location was minimized. The best PPP-performance was achieved when F0 was completely eliminated from the landmark calculation (i.e., when PPP was 0:10).

We subsequently excluded the unaccented syllables from the dataset and run a model on accented syllables only, examining the role of the pitch accent type (High or Low, see Table 2). Again, the model yielded only one main effect of the pitch-to-power ratio (β = 0.0007, *STE* = 0.0002, *t* = 3.11, *p* <.01), indicating that the linear relationship between asynchronies and pitch-to-power ratios also holds in the reduced dataset. Pitch accent type did not improve model fit either as a main effect (β = 0.003, *STE* = 0.002, *t* = 1.42, *p* =.14) or as an interaction term (β = 0.0008, *STE* = 0.00002, *t* = 1.57, *p* =.12).

Finally, we fitted a model to all syllables specified as either carrying a high or a low tone, collapsing across prominence-lending and edge-delimiting tones.[b] The best-fit model retained the interaction of pitch-to-power ratios and tonal specifications (β = 0.0001, *STE* = 0.00004, *t* = 3.37, *p* <.001), which is plotted

[b] We could not examine the role of boundary strength or phrase-delimiting tones on their own because of a relatively small number of phrasal and intonational boundary tones and an imbalanced count of H and L tones (see Table 2).

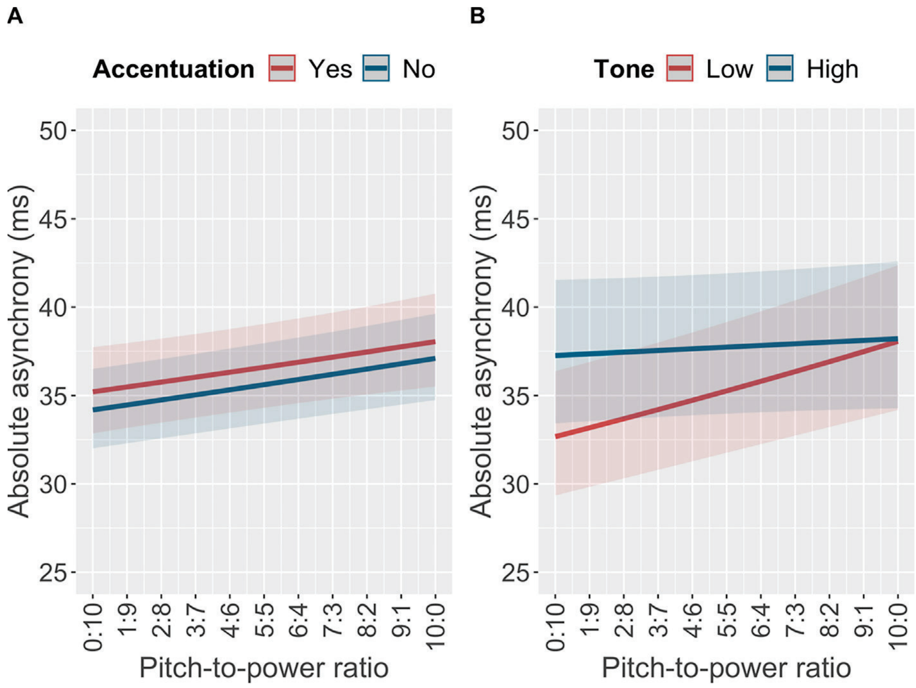

Fig. 5. Linear relationship between predicted absolute asynchronies (in ms) and implemented pitch-to-power ratios: (A) for all syllables of the speech materials, split by the presence of phrase-level accentuation, and (B) for all tone-bearing syllables regardless of their function (prominence-lending or edge-delimiting). Note that a larger asynchrony indicates that the landmark was further away from the adjacent tapping peak.

in Fig. 5B. The interaction indicates that low-pitched syllables display a pattern similar to the one established for the complete dataset. That is, syllables carrying L*, H + L*, L*+H or %L, L-, L-L% show lower the asynchronies with those PPP landmarks that have a smaller contribution of F0 (β = 0.01, *STE* = 0.002, *t* = 5.57, *p* <.001). In contrast, the prediction of the tap location did not improve in high-pitched syllables carrying H*, L + H* or %H, H-, regardless of the variation in the relative contribution of F0 (β = 0.0005, *STE* = 0.00003, *t* = 0.00, *p* =.38). Taken together, these results indicate that F0 is not helpful in predicting the location of taps and suggests that F0-variability of the acoustic signal has little influence on rhythm perception as measured by the sensorimotor synchronization paradigm of the present study.

3.2. IMF minima and maxima

Two separate linear mixed-effects models were performed to test the ability of minima vs. maxima in IMF1 (the first model, representing syllable-level periodicities) and IMF2 (the second model, representing inter-stress periodicities) to predict asynchronies between these landmarks and the nearby tapping peaks. The results of both models showed that the landmarks derived from IMF-maxima outperformed those derived from IMF-minima (IMF1: β = 0.071, *STE* = 0.023, *t* = 3.10, *p* =.002; IMF2: β = 0.066, *STE* = 0.029, *t* = 2.29, *p* =.022, see Fig. 6). Asynchronies measured with the local maxima were significantly smaller than those measured with the local minima, indicating higher predictive power of attracting a tap for the former.

3.3. Asynchronies with automatically derived vs. manually annotated landmarks

A linear mixed-effects model was conducted to compare absolute asynchronies of taps produced in the proximity of all acoustic landmarks. Only the best-fit predictors of PPP and IMF entered the model (i.e., the 0:10 pitch-to-power ratio for PPP and the maxima of IMF1 and IMF2). The predictor *landmark* had five levels (nucleus, *maxD*, PPP, IMF1, and IMF2). The asynchrony model yielded a significant effect of landmark (*SS* = 40.52, *MS* = 10.13, *F*(4) = 9.24, *p* <.001). Subsequent planned factor-level comparisons with the α-level adjusted to 0.0125 (i.e. 0.05/4) showed that the shortest asynchronies were indeed produced with the manually defined nucleus onsets. These landmarks differed significantly from the longer asynchronies measured with PPP (β = 0.103, *STE* = 0.021, *t* = 5.01, *p* <.001), IMF-1 (β = 0.079, *STE* = 0.022, *t* = 3.56, *p* <.001), and IMF-2 (β = 0.093, *STE* = 0.026, *t* = 3.56, *p* <.001). The algorithm-derived landmark *maxD* showed numerically but not significantly longer asynchronies compared to the manually defined nucleus onsets (β = 0.016, *STE* = 0.020, *t* = 0.81, *p* =.42). The results of these comparisons are plotted in Fig. 7. It illustrates the main finding that nucleus onsets were the most likely anchors of sensorimotor synchronization in the present study while PPP landmarks were least likely to anchor stable synchronized movement.

3.4. Probabilities of tapping peaks around automatically derived vs. manually annotated landmarks

A logistic mixed-effects regression was performed to compare the ability of the best performing landmarks to adequately capture the presence of peaks in the tapping distribution. Setting the maximum probability to one tapping peak per syllable, we identified whether or not a tapping peak was located within the predefined window of ± 120 ms around a landmark within each syllable (Repp, 2004, Rathcke et al., 2021), coding this dependent variable as 1 (a tapping peak captured) or 0 (a tap-

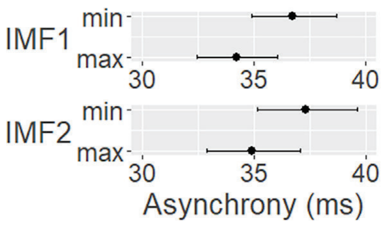

Fig. 6. Estimated asynchronies predicted by maxima vs. minima of IMF1 (upper graph) and IMF2 (lower graph). Note that a larger asynchrony indicates that the landmark was further away from the adjacent tapping peak.

ping peak missed). Again, the predictor *landmark* had five levels (nucleus, *maxD*, PPP, IMF1, and IMF2).

The tapping probability model yielded a significant effect of *landmark* (*AIC* = 37523, χ^2 = 8001.6, *p* <.001). Subsequent planned factor-level comparisons with the α-level adjusted to 0.0125 (i.e. 0.05/4) showed that manually annotated nucleus onsets captured significantly more tapping peaks than three automatically derived landmarks, including PPP (β = −1.23, *STE* = 0.049, *t* = −24.77, *p* <.001), IMF-1 (β = −2.27, *STE* = 0.049, *t* = −45.95, *p* <.001), and IMF-2 (β = −3.27, *STE* = 0.052, *t* = −62.74, *p* <.001). Again, the algorithm-derived landmark *maxD* showed a slight numerical tendency toward capturing fewer tapping peaks than nucleus onsets but the difference between these two landmarks was not significant (β = −0.020, *STE* = 0.055, *t* = −0.36, *p* =.72). The results of the tapping probability model are shown in Fig. 8. Here, the main finding indicates that nucleus onsets captured most tapping peaks of the distribution while IMF-2 landmarks captured the least number of tapping peaks. The latter result is not surprising given that IMF-2 is meant to represent stress-level periodicities while tapping peaks prevail at the syllable level. In this regard, the IMF-1 result is rather surprising as it is meant to represent syllable-level periodicities and yet captures rhythmic entrainment at the syllable level quite poorly, missing more than 50% of syllables and tapping peaks.

3.5. Tapping variability with automatically derived vs. manually annotated landmarks

Finally, the variability of the sensorimotor synchronization behavior was compared across the five landmarks of the study (cf. Leonard and Cummins, 2011). The dependent variable was the standard error of signed asynchronies. As above, the predictor of interest *landmark* had five levels (nucleus, *maxD*, PPP, IMF1, and IMF2). A linear mixed-effects model yielded a significant effect of landmark (*SS* = 25.28, *MS* = 6.32, *F*(4) = 46.44, *p* <.001). Compared to the manually defined nucleus onset, significantly higher levels of sensorimotor variability were measured for PPP (β = 0.009, *STE* = 0.002, *t* = 4.48, *p* <.001), IMF-1 (β = 0.020, *STE* = 0.002, *t* = 9.65, *p* <.001), and IMF-2 (β = 0.023, *STE* = 0.002, *t* = 10.64, *p* <.001) while *maxD* displayed a slight but statistically negligible increase in variability levels (β = 0.003, *STE* = 0.002, *t* = 1.21, *p* =.23). The results of these comparisons are plotted in Fig. 9.

3.6. Consonantal vs. vocalic nuclei

Finally, we examined if the best-performing landmarks – nucleus onset and *maxD* – differed in their ability to predict

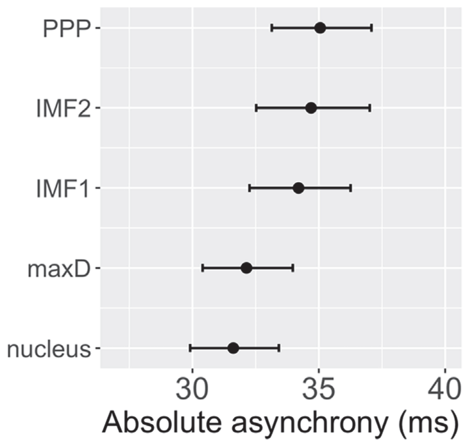

Fig. 7. Comparison between the acoustic landmarks under investigation in their ability to predict the tap location. Note that a larger asynchrony indicates that a tap was further away from the corresponding landmark.

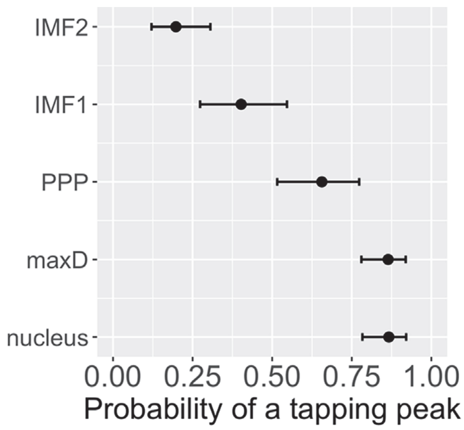

Fig. 8. Comparison of the five landmarks in the estimated probability of capturing a tapping peak within a syllable. The lower the probability, the more tapping peaks are missed by a given landmark.

the tap location, attract a tap, or reduce sensorimotor variability in syllables with vocalic vs. consonantal nuclei. These comparisons only included syllables without lexical stress. The twenty sentences of the present study comprised 120 unstressed syllables in total, 29 of them contained consonantal nuclei and 91 of them contained vocalic nuclei (see Table 1 for the full list of sentences). Given that the total number of syllables with consonantal nuclei is relatively small, we can expect all measurements to have larger confidence intervals and less precise estimates, due to a higher weight given to individual outliers.

3.6.1. Absolute asynchrony: Vocalic vs. consonantal nuclei in maxD and nucleus onset

The results of the linear mixed-effects model showed that absolute asynchronies measured in syllables with vocalic nuclei were not statistically different from those measured in consonantal nuclei, either for the nucleus onset (β = −0.085, *STE* = 0.049, *t* = −1.71, *p* =.088) or for *maxD* (β = −0.064, *STE* = 0.050, *t* = −1.28, *p* =.20). These comparisons are displayed in Fig. 10.

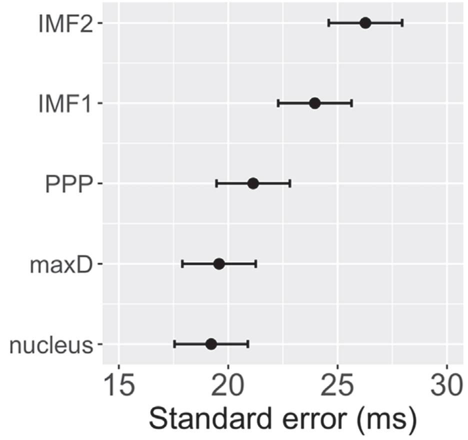

Fig. 9. Comparison of sensorimotor variability (standard error, in ms) measured around the five landmarks under investigation. The lower the variability, the more stable synchronization with the corresponding landmark.

3.6.2. Tapping probabilities: Vocalic vs. Consonantal nuclei in maxD and nucleus onset

The results of the logistic mixed-effects regression showed that the probability of a tap was slightly higher for syllables with vocalic than consonantal nuclei, both for the nucleus onset (β = 0.465, *STE* = 0.113, *z* = 4.11, *p* <.001) and *maxD* (β = 0.553, *STE* = 0.118, *z* = 4.68, *p* <.001, see Fig. 11). However, both types of syllables displayed a high probability (P > 0.75) of being the target of synchronization.

3.6.3. Tapping variability: Vocalic vs. consonantal nuclei in maxD and nucleus onset

The results of the linear mixed-effects model fit to standard error of synchronization showed that tapping was slightly but significantly more variable in syllables with consonantal than vocalic nuclei, measured either with manually identified nucleus onsets (β = 0.146, *STE* = 0.063, *t* = 2.32, *p* <.05) or with automatically derived *maxD* (β = 0.119, *STE* = 0.059, *t* = 2.01, *p* <.05, see Fig. 12).

3.6.4. Nucleus onset vs. maxD in vocalic nuclei

The twenty sentences of the present study contained 167 vocalic nuclei in total. The results of the linear mixed-effects model confirmed that *maxD* performed no different to the nucleus onset (β = 0.016, *STE* = 0.022, *t* = 0.73, *p* =.47) in terms of predicting tap locations. A permutation test (based on 5,000 permutations) was additionally conducted using R-package *predictmeans* (Luo et al., 2020). In each permutation, conditions were randomly assigned to the data points, and a statistical test was run to acquire the *F*-score. From the 5,000 permutations conducted, 95% of all tests had a *F*-score below 3.85 (Fig. 12), indicating that a F-score above the cut-off value of 3.85 was needed to support the claim that the two conditions (*onset* vs. *maxD*) were indeed significantly different from each other (the cut-off is similar to *p* <.05). However, the observed *F*-score was far away from the calculated cut-off value, yielding a non-significant result (*F* = 0.57, *p* =.44). This result further confirmed the absence of a difference between asynchronies measured with *maxD* vs. nucleus onsets (Fig. 13).

3.6.5. Nucleus onset vs. maxD in consonantal nuclei

The twenty sentences of the present study contained 29 consonantal nuclei in total, i.e., this analysis was based on a smaller subset of the data. Here, the results of the linear mixed-effects models again confirmed the main finding that *maxD* performed similar to a nucleus onset, even when a sonorant consonant constituted syllable nucleus (β = 0.028, *STE* = 0.055, *t* = 0.52, *p* =.60). An additional permutation test (based on 5,000 permutations) further confirmed the finding (observed *F* = 0.22; *p* =.64). Once again, the observed *F*-score did not pass the extreme value (95th percentile = 3.93) in the permutation test comparing asynchronies with *maxD* vs. nucleus onsets. A graph of the corresponding F-score distribution is not shown here, given its close similarity to Fig. 13 shown before.

4. Discussion

The overarching goal of the present study was to identify the foundational unit of rhythm (the beat) in varied, naturalistic speech materials by means of a sensorimotor synchronization paradigm. The task required study participants to synchronize with what they felt to be the beat of natural spoken sentences played to them on a short loop. The sentences included syllables with vocalic as well as consonantal nuclei and varied in syllable number and duration. We tested several linguistic and acoustic landmarks as potential anchors of the rhythmic movement and further examined the contribution of pitch to the anchoring potential of a landmark. The results of the study make four innovative contributions to the study of speech rhythm, as outlined below.

4.1. The role of pitch in sensorimotor synchronization with natural speech

The first aim of the present study was to investigate the role of pitch in anchoring synchronized movement with natural speech. The results obtained with the PPP-algorithm provided evidence for a linear relationship between pitch-to-power ratios and tapping asynchrony, namely the lower the weighting of F0, the closer the tap to the resulting landmark. The result also holds if the dataset is reduced to phonological tone-bearing units only. This relationship suggests that pitch does not play any role in establishing a reliable anchor of sensorimotor synchronization with speech, at least as far as English is concerned.

This result is quite different from what might be expected based on empirical evidence that pointing gestures in a picture-naming task tend to align with F0 peaks of pitch-accented syllables (Esteve-Gibert & Prieto, 2013; Leonard & Cummins, 2011). There is of course a number of major differences between the two types of studies that might explain the different findings, including differences in speech materials (broad focus sentences in the present study vs. contrastive focus sentences in the study by Esteve-Gibert & Prieto, 2013), the task itself (synchronizing with the beat of an externally-presented looped sentence vs. pointing out one salient word in a self-produced sentence), biomechanics of finger movement (tapping an index finger on a hard surface vs. raising an index finger high in the air) etc. Last but not least, the

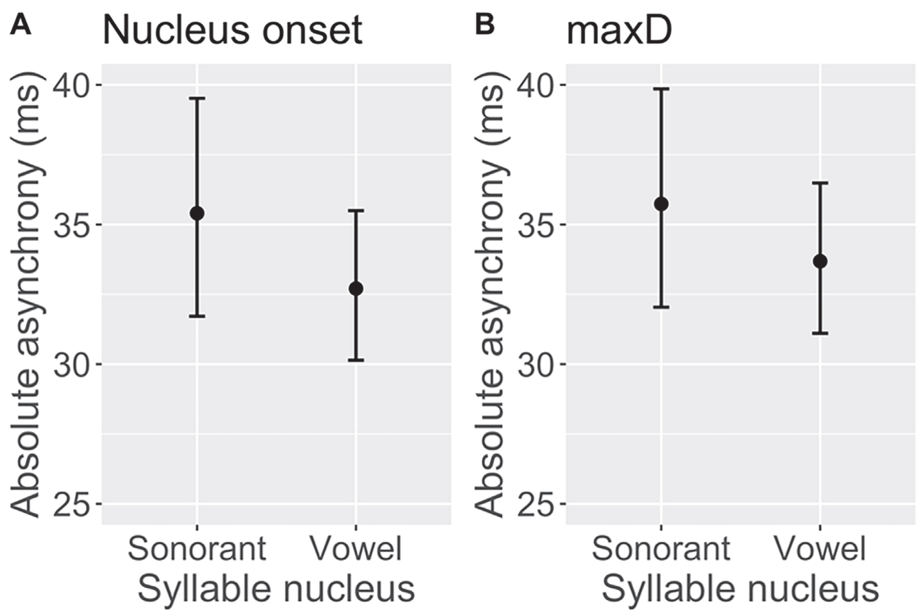

Fig. 10. Estimated absolute asynchronies of sensorimotor synchronization with weak syllables containing consonantal vs. vocalic nuclei. Results for the manually defined onset landmarks are given on the left. Results for the algorithm-generated *maxD* are shown on the right.

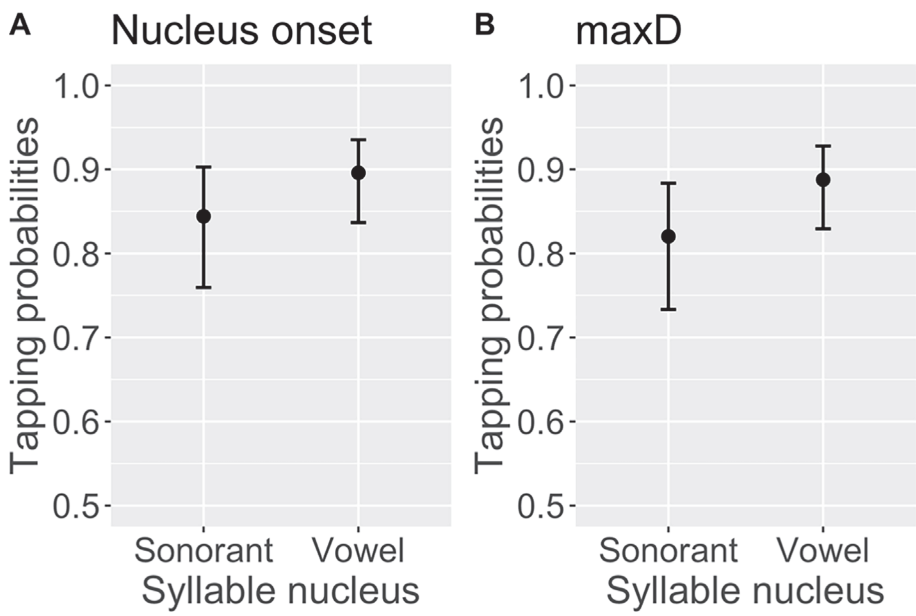

Fig. 11. Estimated tapping probabilities of sensorimotor synchronization with weak syllables containing consonantal vs. vocalic nuclei. A local tapping peak must fall within ± 120 ms window around a nucleus onset or a *maxD* landmark to be included. Results for manually defined onsets are given on the left (A). Results for algorithm-generated *maxD* are shown on the right (B).

procedure of deriving the acoustic landmarks in sentences differs between these studies, making a direct comparison of the synchronization results challenging. In addition, the beat is but one aspect of the complex phenomenon that is rhythm. It may be possible that pitch plays a role in sensorimotor synchronization with higher levels of rhythmic structure in natural speech, specifically meter (Hannon et al., 2004; Large & Snyder, 2009). It is also possible that pitch is more substantially involved in guiding rhythmic attention in languages other than English, e.g. in French that relies on distinct pitch movements for delimiting the phonological unit of an accentual phrase (Jun & Fougeron, 2002), or German that has a more transparent mapping between pitch and prominence than some other languages (Riesberg et al., 2020). At the same time, it is known that the relationship between pitch and perceived rhythm is quite complex (Thomassen, 1982). Recent evidence further shows that native speakers of tonal languages outperform native speakers of non-tonal languages on musical pitch tasks but perform significantly worse on musical beat tasks, suggesting a perceptual trade-off between strong auditory skills of pitch vs. rhythm processing (Liu et al., 2023; Zhang et al., 2020). Future work will thus benefit from cross-linguistic comparisons of the role of pitch in rhythm perception.

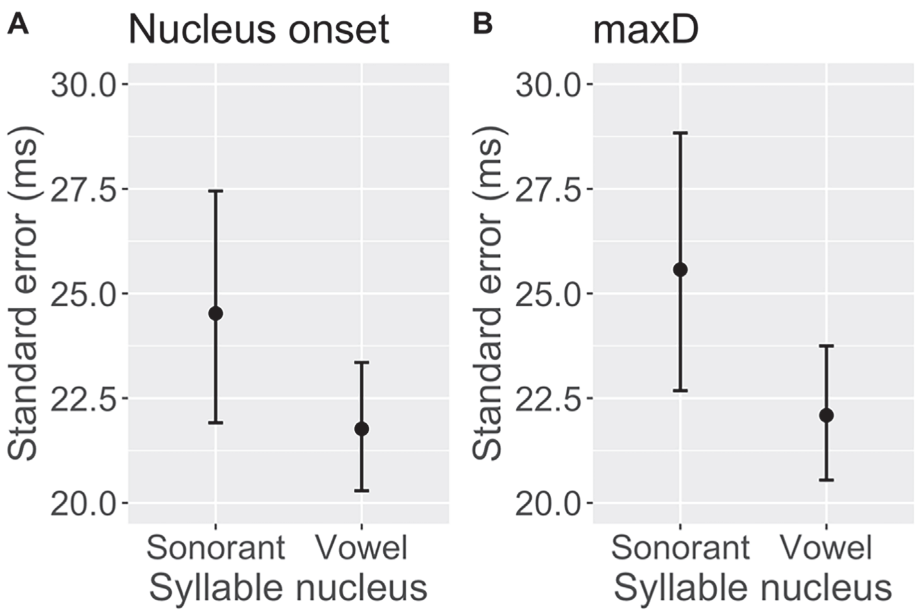

Fig. 12. Estimated tapping variability of sensorimotor synchronization with weak syllables containing consonantal vs. vocalic nuclei, for (A) the manually defined onsets and (B) the algorithm-generated *maxD*.

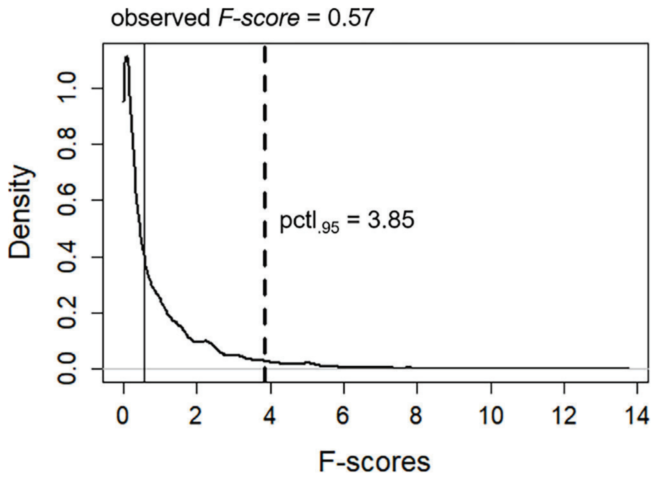

Fig. 13. The *F*-score distribution of the 5,000 permutation tests comparing asynchronies with nucleus onsets vs. *maxD*. The observed *F*-score was not larger than the 95th percentile ($\alpha = 0.05$).

4.2. Signal decomposition into multiple timescales

The second aim of the present study was to investigate if sensorimotor synchronization with natural speech is best predicted by specific timepoints identified in the periodic components of decomposed speech signals. This type of signal processing unveils underlying periodicities at inter-syllabic and inter-stress timescales that can display meaningful cross-linguistic differences in rhythmic structure (Tilsen & Arvaniti, 2013). A pilot study (Lin & Rathcke, 2020) examined if local maxima of IMFs derived by the empirical mode decomposition (Huang et al., 1998) would serve as tap anchors, but did not provide empirical support for the idea. The present study extended the database for the analyses and further included local minima of intrinsic mode functions as the onsets of periodicities on the inter-syllable and inter-stress timescales.

However, IMF-minima on both timescales performed worse than the corresponding IMF-maxima as far as the prediction of tap location was concerned. Moreover, even the better-performing IMF-maxima did not deal well with capturing the sensorimotor synchronization behavior of study participants. They performed notably worse on predicting asynchronies, reducing sensorimotor variability, and capturing synchronization peaks than other landmarks of the study (and significantly worse than nucleus onsets). Long-scale signal decomposition methods may excel at tracking temporal periodicities on several, concomitantly present scales (Tilsen & Arvaniti, 2013), but they seem to lack temporal precision at the local level of rhythmic units that is tapped into by the sensorimotor synchronization paradigm.

4.3. Landmarks anchoring sensorimotor synchronization

The third aim of the present study was to compare synchronization patterns of algorithm-generated acoustic landmarks to manually defined nucleus onsets, and to test if the results apply to vocalic and consonantal nuclei in a similar way.

We found that only *maxD* performed as well as the nucleus onset which is fully in line with our previous work on a smaller set of linguistic materials (Lin & Rathcke, 2020). Again, we found a non-significant, slight numerical difference between asynchronies with *maxD* vs. the nucleus onset, that indicated a slight superiority of the onset over the *maxD* landmark (Lin & Rathcke, 2020; Rathcke, Lin, et al., 2021). We further confirmed that the two landmarks produced comparable results when capturing tapping probabilities and reducing tapping variability.

Given that *maxD* is automatically derived by an algorithm that requires no more than the annotation of syllable boundaries as input, *maxD* can be recommended for studies that rely on large stimulus pools and generally when manual annotation by a trained phonetician is unavailable or difficult due to the segmental composition of stimuli (Turk et al., 2012). The method can be further automatized by implementing algorithm-driven annotations of syllable boundaries instead of the manual annotations implemented here, making the *maxD* method even less time-consuming. The discussion of whether or not *maxD* indeed represents the P-centre as has been suggested by Šturm & Volín (2016) is beyond the scope of the study. However, the present finding suggests that the local increase in the acoustic energy does not only reflect onsets of rhythmically relevant events in short mono- or bisyllabic words (that were tested in the original study with Czech materials), but also captures rhythmic structure in longer, natural sentences. Moreover, the algorithm can be applied to languages other than Czech and generally appears to be applicable cross-linguistically, at least in those languages that permit relatively complex onset phonotactics such as English and Czech (Kessler & Treiman, 1997; Selkirk, 1982). An alternative, potentially fruitful avenue to follow in future research is to adopt automated routines in the annotation of nucleus onsets (cf. Adi et al., 2016).

The relatively poor performance of PPP (even after the elimination of pitch information) might be reflective of the fact that local energy maxima do not seem to serve as reliable tap attractors (Lin & Rathcke, 2020; Rathcke et al., 2019) and do not play a role in capturing P-center locations (Marcus, 1981; Morton et al., 1976; Šturm & Volín, 2016). Even though periodic power is not equal to pure intensity, the two aspects of acoustic signals are still closely related (Albert et al., 2018). The PPP-results are therefore not surprising.

In general, the landmarks derived from IMFs performed worse than the manually identified nucleus onsets (and all other landmarks for that matter). These findings might be at least partly due to the fact that IMFs are the only landmarks of the present study that can be generated without any manual input or supervision. The other automatically derived landmarks (i.e., *maxD* and PPP) require manually annotated syllable boundaries that guide the algorithm performance. It seems that, despite the temporal realignment of the energy peaks between the derived IMFs and the original signals that we performed manually for the purpose of this study, the decompositional approach might still not adequately capture the local timing relations in natural speech. Since automated data processing brings a methodological advantage to many workflows (e.g., Sonderegger et al., 2020), there might be further algorithms to explore in the study of rhythmic synchronization with natural speech. For example, the continuous wavelet transform approach to tracing the prosodic hierarchy (Suni et al., 2017) may be promising in this context.

Finally, we did not establish different synchronization anchors for syllables with vocalic vs. consonantal nuclei as the general patterns of synchronization remained comparable across the two types of syllables. Participants synchronized with both types of weak syllables, even if they tapped with the consonantal nuclei slightly less often and slightly less consistently. This important finding implies that it is not the sole presence of vowels that guides rhythmic attention in natural speech but the presence of a linguistic unit such as syllable that constitutes a low level of prosodic hierarchy in natural speech (Ding et al., 2015; Hayes, 1989).

4.4. The domain of rhythmic attention in natural speech

The results of the present study support our hypothesis that rhythmic attention in speech is guided by both the linguistic knowledge and the acoustic properties of speech signals (cf. Brown, Tanenhaus, & Dilley, 2021). Absolute asynchronies did not differ between consonantal and vocalic nuclei, and the probability of a tap being recorded in a syllable containing a consonantal nucleus, even though slightly lower than in a syllable containing a vowel nucleus, did not equal zero. Tapping variability was slightly increased in syllables with sonorant consonants, though the difference was again very subtle and possibly due to a lower number of syllables containing consonantal nuclei as compared to more frequent vocalic nuclei. Taken together, these results indicate that targets of sensorimotor synchronization are not vowels *per se* but rather syllabic nuclei. That is, rhythmic attention seems to be generally locked onto the nucleus of a syllable, and not just a vowel. Relatedly, previous research has shown that the nucleus type (vowel or consonant) does not influence the preceding or the following consonant duration (Beňuš, 2011), suggesting that a stable temporal structure within a syllable can be formed regardless of the nucleus type.

Syllable nucleus identification and segmentation is crucial for speech rhythm perception as it represents the syllable rate of speech (Zhang & Glass, 2009). Distal rate effects encoded by the syllable rate are important for the perceptual normalization of timing cues that guide word segmentation (Dilley & McAuley, 2008; Heffner et al., 2013). Features of the syllable-level rhythm are believed to be crucial for rhythm perception in speech (e.g., Barbosa, 2007; Leong & Goswami, 2014; O'Dell & Nieminen, 1999; Saltzman, Nam, Krivokapic, & Goldstein, 2008) and to account for cross-linguistic differences (Tilsen & Arvaniti, 2013). According to previous research, syllabic oscillations derived from amplitude envelopes of speech signals can capture key rhythmic features characteristic of different languages or different speech styles (Tilsen & Arvaniti, 2013). Moreover, syllable rate has been identified as an important source of acoustic information for neural entrainment in the theta band that accompanies word perception (Giraud & Poeppel, 2012; Peelle, Gross, & Davis, 2013).

The present finding corroborates proposals that recognize the cross-linguistically foundational role of syllables in language production (Iskarous & Pouplier, 2022) and language processing (Brown, Tanenhaus, & Dilley, 2021), though it also calls into question assumptions that the systemic relevance of syllable units is due to the local rises in acoustic intensity, that are associated specifically with vowel onsets (Doelling et al., 2014; Oganian & Chang, 2019). Cross-linguistic diversity in syllable structure encompasses syllabic consonants alongside vowels. There might be the reason why syllables with consonantal nuclei are likely to be easily recognizable as a unit of speech (Reddy et al., 2013), thus serving as a good target of sensorimotor synchronization. They display a relatively high sonority level, with the intensity of the resulting syllable being also relatively high (Parker, 2002). Future research will ultimately need to address the question of whether or not voiceless nuclei with relatively low sonority and low level of syllabic energy (e.g., as found in Tashlhiyt Berber, Ridouane,

2002) can anchor rhythmic attention in a similar way, in order to further corroborate the present conclusions.

4.5. Methodological considerations

In contrast to naturalistic paradigms that study pitch-gesture synchronisation in conversation (Esteve-Gibert & Prieto, 2013; Leonard & Cummins, 2011), synchronization with looped linguistic stimuli is certainly a more laboratory task, akin to word repetition paradigms employed in articulatory phonology (e.g., Pouplier, 2008; Pouplier & Goldstein, 2010). In the sensorimotor synchronization task, sentence repetition is the core prerequisite for an individually stable synchronization performance with natural speech, given that participants require at least two repetitions of a sentence prior to developing "the feeling of the beat" and approximately five repetitions prior to establishing a stable synchronization pattern (Rathcke, Lin, et al., 2021). Yet unlike the study of sensorimotor synchronization with a repeated metronome beat, the goal of present research is not to capture time series of taps in order to shed light on the physiological limits of human motor skills or clinical patterns of perceptual timing abilities (Dalla Bella et al. 2017; Repp, 2005, 2006, 2010; Repp & Su, 2013). Rather, it seeks to understand the nature of rhythmic attention in spoken language by means of a behavioural task and as such, it does not crucially hinge on a precise estimation of temporal synchronisation error over the course of sentence repetitions.

Without doubts, SMS with real, natural speech is more challenging than SMS with a metronome (cf. Dalla Bella et al., 2013), potentially leading to a higher level of individual uncertainty before a stable performance can be achieved and maintained (Rathcke, Lin, et al., 2021). Individual participants might lag behind SMS-targets at first and start anticipating them later in the loop, or they might anticipate at first and stabilize with a 0 ms synchronization error later. What therefore matters less in SMS with real speech is the exact magnitude of synchronization error as it varies in this complex task. What, however, matters more is the likelihood of certain speech events to reliably attract SMS throughout sentence repetitions as well as the temporal relation between tapping probabilities and local attractors. This kind of information is difficult to obtain from a time-series approach to SMS as it is usually deployed in studies of SMS with a metronome. In the worst case, a time-series based analysis can result in conceptually misleading representations, especially when individual SMS-behaviour changes from lagging to anticipating during repetitions. In such cases, the synchronization error derived from the corresponding time series is likely to average at 0 ms, misrepresenting both the error and the underlying SMS-behaviour. A relatively high level of uncertainty and related SMS-variability is the main reason why aggregation and probability density estimation are key to adequately capturing SMS with natural speech (Rathcke, Lin, et al., 2021). Kernel-smoothed distributions help to deal with the trial-by-trial variability in that they reflect the most likely locations of taps across repetitions. A similar (adaptive kernel-density estimation) approach has been used in related SMS-studies, to address the issue of large individual variability and to provide a reliable representation of head movements measured during synchronization with complex auditory signals such as music (Dotov et al., 2021). Note, however, that the asynchrony value derived by this method may deviate from the actual asynchrony measured with a specific landmark on a specific trial. If the goals of a given SMS-study demand direct and precise assessment of asynchrony with a local landmark, a viable alternative may involve calculations of asynchrony in one (e.g., the last as the most stable) cycle of sentence repetitions. While this approach loses the information about the relative likelihood of the given landmark to attract SMS, it provides temporal precision of measuring local SMS-behaviour.

The ability to synchronize and "the feeling of the beat" do not appear to be moderated by the perceptual experience of speech shifting to song that may sometimes happen during repetitions of spoken phrases (Deutsch, 1995). Our previous work compared sensorimotor synchronization with sentences that had transformed from speech to song very frequently or very rarely (Rathcke, Lin, et al., 2021). We found that synchronization accuracy was comparable across high- and low-transforming sentences, with the only difference that the overall number of recorded taps was slightly higher in the high-transforming than in the low-transforming sentences. The different frequencies of sensorimotor synchronization with the two types of sentences are unlikely to arise from their differential ability to cause the transformation from speech to song. Instead, it may be related to a high level of the overall signal sonority that was the main acoustic feature characteristic of the high-transforming sentence set (Rathcke, Falk, et al., 2021).

Linguistic rhythm is a distal prosodic feature (cf. Baese-Berk et al., 2019; Morrill et al., 2015) and as such, requires exposure on long timescales. In this sense, looping creates the experience of a prolonged exposure to the rhythmic structure of speech (while bypassing other mechanisms of sentence processing, cf. Rathcke, Falk, et al., 2021). Similar approaches to the study of linguistic rhythm recruit very special cases of language use, e.g. mantra or chant (Cummins, 2012) or experiment with short, regularized sentences (Lidji et al., 2011). In comparison to such paradigms, looping employed in the present study has the advantage of being applicable to any type of speech, including spontaneous productions which are often considered less rhythmic or more difficult to capture rhythmically (Kohler, 2009).

5. Conclusion

The present study employed a movement-based paradigm to study rhythm perception via entrainment with spoken sentences. The results demonstrated that people tended to synchronize with onsets of syllable nuclei (both vocalic and consonantal). A derivative of the acoustic energy contour - *maxD* - that describes the fastest increase of periodic energy around the nucleus onset can also be viewed as a representative target of sensorimotor synchronization with speech. The acoustically determined *maxD* has the advantage of being derived by an automated routine instead of a manual inspection of speech signals. Moreover, the present study shows that F0-information appears to have little effect on anchoring sensorimotor synchronization in a finger-tapping task. These results are limited to the lower levels of prosodic hierarchy which we assume to reflect the beat in speech, and they are also limited to English (and possibly other, prosodically similar languages). Future research ought to address if the conclusions of the present study also hold for synchronization with higher levels of rhythmic structure and across prosodically diverse languages.

CRediT authorship contribution statement

Tamara Rathcke: Conceptualization, Funding acquisition, Formal analysis, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Chia-Yuan Lin:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The study was funded by the Leverhulme Trust (RPG-2017-306 grant to the first author). The authors would like to thank Simone Dalla Bella and Simone Falk for valuable discussions of the pre-processing involved in the analyses of sensorimotor synchronization data, the team at BRAMS in Montreal for the support and training provided to the second author during his stay at BRAMS that was funded by the British Psychological Society. The authors are further grateful to Aviad Albert and Francesco Cangemi for their generosity in sharing various scripts, Sam Tilsen for his support with the EMD analyses, Pavel Šturm for sharing the *maxD* script. The authors' special thanks go to Niamh Molloy who assisted with the data collection.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh University Press.

Adi, Y., Keshet, J., Cibelli, E., Gustafson, E., Clopper, C., & Goldrick, M. (2016). Automatic measurement of vowel duration via structured prediction. *The Journal of the Acoustical Society of America, 140*(6), 4517–4527.

Albert, A., Cangemi, F., & Grice, M. (2018). Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. In *Proc. 9th International Conference on Speech Prosody 2018* (pp. 804–808). https://doi.org/10.21437/SpeechProsody.2018-162.

Allen, G. D. (1972). The location of rhythmic stress beats in English : An experimental study I. *Language and Speech, 15*(1), 72–100. https://doi.org/10.1177/002383097201500208.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica, 66*(1–2), 46–63. https://doi.org/10.1159/000208930.

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics, 40*(3), 351–373. https://doi.org/10.1016/j.wocn.2012.02.003.

Aschersleben, G. (2002). Temporal control of movements in sensorimotor synchronization. *Brain and Cognition, 48*(1), 66–79. https://doi.org/10.1006/brcg.2001.1304.

Baese-Berk, M. M., Dilley, L. C., Henry, M. J., Vinke, L., & Banzina, E. (2019). Not just a function of function words: Distal speech rate influences perception of prosodically weak syllables. *Attention, Perception, and Psychophysics, 81*(2). https://doi.org/10.3758/s13414-018-1626-4.

Barbosa, P. A. (2007). From syntax to acoustic duration: A dynamical model of speech rhythm production. *Speech Communication, 49*(9), 725–742. https://doi.org/10.1016/j.specom.2007.04.013.

Beňuš, Š. (2011). Control of phonemic length contrast and speech rate in vocalic and consonantal syllable nuclei. *The Journal of the Acoustical Society of America, 130*(4), 2116–2127. https://doi.org/10.1121/1.3624824.

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer [Computer program]*. Version 6.0.43. http://www.praat.org.

Brown, M., Tanenhaus, M. K., & Dilley, L. (2021). Syllable inference as a mechanism for spoken language understanding. *Topics in Cognitive Science, 13*(2), 351–398. https://doi.org/10.1111/tops.12529.

Cakewalk by Bandlab [computer program]. (2019). Cakewalk Inc. https://www.cakewalk.com/.

Cangemi, F., Albert, A., & Grice, M. (2019). Modelling intonation: Beyond segments and tonal targets. In *Proceedings of the 19th International Congress of Phonetic Sciences, August* (pp. 572–576) https://www.researchgate.net/publication/335096495.

Chen, J. L., Penhune, V. B., & Zatorre, R. J. (2008). Listening to musical rhythms recruits motor regions of the brain. *Cerebral Cortex, 18*(12), 2844–2854. https://doi.org/10.1093/cercor/bhn042.

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

Cummins, F. (2012). Looking for rhythm in speech. *Empirical Musicology Review, 7*(1), 28–35. https://doi.org/10.18061/1811/52976.

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

Cutler, A. (2015). Lexical stress in English pronunciation. *The Handbook of English Pronunciation*, 106–124.

Dalla Bella, S., Białuńska, A., & Sowiński, J. (2013). Why movement Is captured by music, but less by speech: Role of temporal regularity. *PLoS ONE, 8*(8), 1–16. https://doi.org/10.1371/journal.pone.0071945.

Dalla Bella, S., Farrugia, N., Benoit, C. E., Begel, V., Verga, L., Harding, E., & Kotz, S. A. (2017). BAASTA: Battery for the assessment of auditory sensorimotor and timing abilities. *Behavior Research Methods, 49*(3), 1128–1145. https://doi.org/10.3758/s13428-016-0773-6.

Deutsch, D. (1995). *Musical illusions and paradoxes*. CD: Philomel Records.

Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language, 59*(3), 294–311. https://doi.org/10.1016/j.jml.2008.06.006.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2015). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience, 19*(1), 158–164. https://doi.org/10.1038/nn.4186.Cortical.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage, 85*, 761–768. https://doi.org/10.1016/j.neuroimage.2013.06.035.

Dotov, D., Bosnyak, D., & Trainor, L. J. (2021). Collective music listening: Movement energy is enhanced by groove and visual social cues. *Quarterly Journal of Experimental Psychology, 74*(6). https://doi.org/10.1177/1747021821991793.

Drake, C., & Palmer, C. (1993). Accent Structures in Music Performance. *Music Perception, 10*(3). https://doi.org/10.2307/40285574.

Eerola, T., & Toiviainen, P. (2004). MIDI Toolbox: MATLAB Tools for Music Research. In *University of Jyväskylä: Kopijyvä, Jyväskylä, Finland.* http://www.jyu.fi/musica/miditoolbox/.

Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research, 56*(3), 850–864. https://doi.org/10.1044/1092-4388(2012/12-0049).

Falk, S., Rathcke, T. V., Dalla Bella, S., & Bella, S. D. (2014). When speech sounds like music. *Journal of Experimental Psychology: Human Perception and Performance, 40*(4), 1491–1506. https://doi.org/10.1037/a0036858.

Féry, C. (2003). Markedness, Faithfulness, Vowel Quality and Syllable Structure in French. *Journal of French Language Studies, 13*(2), 247–280. https://doi.org/10.1017/S0959269503001121.

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* (Issue 30, pp. 521–535). Erlbaum.

Fudge, E. C. (1969). Syllables. *Journal of Linguistics, 5*(2), 253–286. https://doi.org/10.1017/S0022226700002267.

Gerhard, D. (2003). *Computationally measurable differences betwen speech and song [PhD thesis]*. Simon Fraser University.

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience, 15*(4), 511–517. https://doi.org/10.1038/nn.3063.

Goswami, U., & Leong, V. (2013). Speech rhythm and temporal structure: Converging perspectives? *Laboratory Phonology, 4*(1), 67–92. https://doi.org/10.1515/lp-2013-0004.

Grahn, J. A., & Rowe, J. B. (2009). Feeling the beat: Premotor and striatal interactions in musicians and nonmusicians during beat perception. *Journal of Neuroscience, 29*(23), 7540–7548. https://doi.org/10.1523/JNEUROSCI.2018-08.2009.

Hannon, E. E., Snyder, J. S., Eerola, T., & Krumhansl, C. L. (2004). The role of melodic and temporal cues in perceiving musical meter. *Journal of Experimental Psychology: Human Perception and Performance, 30*(5), 956. https://doi.org/10.1037/0096-1523.30.5.956.

Hawkins, S. (2014). Situational influences on rhythmicity in speech, music, and their interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1658). https://doi.org/10.1098/rstb.2013.0398.

Hayes, B. (1989). The Prosodic Hierarchy in Meter. In *Rhythm and Meter* (pp. 201–260). Elsevier. https://doi.org/10.1016/b978-0-12-409340-9.50013-9.

Heffner, C. C., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2013). When cues combine: How distal and proximal acoustic cues are integrated in word segmentation. *Language and Cognitive Processes, 28*(9), 1275–1302. https://doi.org/10.1080/01690965.2012.672229.

Howell, P. (1988). Prediction of P-center location from the distribution of energy in the amplitude envelope: I & II. *Perception & Psychophysics, 43*(1), 99. https://doi.org/10.3758/BF03208978.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Snin, H. H., Zheng, Q., ... Liu, H. H. (1998). The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 454*(1971), 903–995. https://doi.org/10.1098/rspa.1998.0193.

Iversen, J. R., Repp, B. H., & Patel, A. D. (2009). Top-down control of rhythm perception modulates early auditory responses. *Annals of the New York Academy of Sciences, 1169*(1), 58–73. https://doi.org/10.1111/j.1749-6632.2009.04579.x.

Iskarous, K., & Pouplier, M. (2022). Advancements of phonetics in the 21st century: A critical appraisal of time and space in Articulatory Phonology. *Journal of Phonetics, 95*, 101195. https://doi.org/10.1016/j.wocn.2022.101195.

Jun, S. A., & Fougeron, C. (2002). Realizations of accentual phrase in French intonation. *Probus, 14*(1), 147–172. https://doi.org/10.1515/prbs.2002.002.

Kahn, D. (2015). Syllable-based generalizations in English phonology. In *Syllable-Based Generalizations in English Phonology*. https://doi.org/10.4324/9781315688121.

Kessler, B., & Treiman, R. (1997). Syllable structure and the distribution of phonemes in english syllables. *Journal of Memory and Language, 37*(3). https://doi.org/10.1006/jmla.1997.2522.

Kohler, K. J. (2008). The perception of prominence patterns. *Phonetica, 65*(4), 257–269. https://doi.org/10.1159/000192795.

Kohler, K. J. (2009). Rhythm in speech and language: A new research paradigm. *Phonetica, 66*(1–2), 29–45. https://doi.org/10.1159/000208929.

Kolassa, J. E. (2020). *An introduction to nonparametric statistics*. CRC Press.

Ladd, D. R. (2008). Intonational Phonology. *Cambridge University Press*. https://doi.org/10.1017/cbo9780511808814.

Ladefoged, P. (1993). *A course in phonetics* (3rd ed.). Harcourt Brace Jovanovich College Publishers.

Ladefoged, P. (1975). *A Course in Phonetics*. NewYork: Harcourt Brace Jovanovich.

Langus, A., Marchetto, E., Bion, R. A. H., & Nespor, M. (2012). Can prosody be used to discover hierarchical structure in continuous speech? *Journal of Memory and Language, 66*(1), 285–306. https://doi.org/10.1016/j.jml.2011.09.004.

Large, E. W., & Palmer, C. (2002). Perceiving temporal regularity in music. *Cognitive Science, 26*(1), 1–37. https://doi.org/10.1016/S0364-0213(01)00057-X.

Large, E. W., & Snyder, J. S. (2009). Pulse and meter as neural resonance. *Annals of the New York Academy of Sciences, 1169*. https://doi.org/10.1111/j.1749-6632.2009.04550.x.

Leonard, T., & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes, 26*(10), 1457–1471. https://doi.org/10.1080/01690965.2010.500218.

Leong, V., & Goswami, U. (2014). Assessment of rhythmic entrainment at multiple timescales in dyslexia: Evidence for disruption to syllable timing. *Hearing Research, 308*, 141–161. https://doi.org/10.1016/j.heares.2013.07.015.

Levin, J. (1985). A metrical theory of syllabicity. In *Ph.D. dissertation, MIT. Lle , C. & M. Prinz* (Vol. 23, Issue 1981).

Lidji, P., Palmer, C., Peretz, I., & Morningstar, M. (2011). Listeners feel the beat: Entrainment to English and French speech rhythms. *Psychonomic Bulletin and Review, 18*(6), 1035–1041. https://doi.org/10.3758/s13423-011-0163-0.

Lin, C.-Y., & Rathcke, T. (2020). How to hit that beat: Testing acoustic anchors of rhythmic movement with speech. In *Proc. 10th International Conference on Speech Prosody 2020, May* (pp. 1–5). https://doi.org/10.21437/SpeechProsody.2020-1.

Lindbolm, B., & Sundberg, J. (2007). The human voice in speech and singing. In T. Rossing (Ed.), *Springer handbook of acoustics* (pp. 669–706). Springer Verlag.

Liu, J., Hilton, C. B., Bergelson, E., & Mehr, C. A. (2023). Language experience predicts music processing in a half-million speakers of fifty-four languages. *Current Biology*. https://doi.org/10.1016/j.cub.2023.03.067.

London, J. (2004). *Hearing in time*. Oxford University Press.

Luo, D., Ganesh, S., & Koolaard, J. (2020). *predictmeans: Calculate Predicted Means for Linear Models*. https://cran.r-project.org/package=predictmeans.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Madison, G. (2014). Sensori-motor synchronisation variability decreases as the number of metrical levels in the stimulus signal increases. *Acta Psychologica, 147*, 10–16. https://doi.org/10.1016/j.actpsy.2013.10.002.

Mahadeva Prasanna, S. R., Sandeep Reddy, B. V., & Krishnamoorthy, P. (2009). Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. *IEEE Transactions on Audio, Speech and Language Processing, 17*(4), 556–565. https://doi.org/10.1109/TASL.2008.2010884.

Manal, K., & Rose, W. (2007). A general solution for the time delay introduced by a low-pass Butterworth digital filter: An application to musculoskeletal modeling. *Journal of Biomechanics, 40*(3), 678–681. https://doi.org/10.1016/j.jbiomech.2006.02.001.

Marchand, Y., Adsett, C. R., & Damper, R. I. (2009). Automatic syllabification in English: A comparison of different algorithms. *Language and Speech, 52*(1), 1–27. https://doi.org/10.1177/0023830908099881.

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

Morillon, B., Schroeder, C. E., & Wyart, V. (2014). Motor contributions to the temporal precision of auditory attention. *Nature Communications, 5*, 1–9. https://doi.org/10.1038/ncomms6255.

Morrill, T., Baese-Berk, M., Heffner, C., & Dilley, L. (2015). Interactions between distal speech rate, linguistic knowledge, and speech environment. *Psychonomic Bulletin and Review, 22*(5). https://doi.org/10.3758/s13423-015-0820-9.

Morrill, T. H., Dilley, L. C., & McAuley, J. D. (2014). Prosodic patterning in distal speech context: Effects of list intonation and f0 downtrend on perception of proximal prosodic structure. *Journal of Phonetics, 46*(1), 68–85. https://doi.org/10.1016/j.wocn.2014.06.001.

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

Niebuhr, O. (2009). F0-based rhythm effects on the perception of local syllable prominence. *Phonetica, 66*(1–2), 95–112. https://doi.org/10.1159/000208933.

Nolan, F., & Jeon, H. S. (2014). Speech rhythm: A metaphor? *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1658), 20130396. https://doi.org/10.1098/rstb.2013.0396.

O'Dell, M. L., & Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 1075–1078).

Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *Science advances, 5*(11), eaay6279. https://doi.org/10.1126/sciadv.aay6279.

Parker, S. G. (2002). *Quantifying the sonority hierarchy*. University of Massachusetts Amherst.

Patel, A. D. (2008). *Music, language and the brain*. Oxford: Oxford University Press.

Phillips-Silver, J., & Trainor, L. J. (2005). Psychology: Feeling the beat: Movement influences infant rhythm perception. *Science, 308*(5727), 1430. https://doi.org/10.1126/science.1110922.

Phillips-Silver, J., & Trainor, L. J. (2007). Hearing what the body feels: Auditory encoding of rhythmic movement. *Cognition, 105*(3), 533–546. https://doi.org/10.1016/j.cognition.2006.11.006.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral cortex, 23*(6), 1378–1387. https://doi.org/10.1093/cercor/bhs118.

Pike, K. L., & Pike, E. V. (1947). Immediate Constituents of Mazateco Syllables. *International Journal of American Linguistics, 13*(2), 78–91. https://doi.org/10.1086/463932.

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics, 17*(3), 175–192. https://doi.org/10.1016/s0095-4470(19)30428-0.

Pouplier, M. (2008). The role of a coda consonant as error trigger in repetition tasks. *Journal of Phonetics, 36*(1). https://doi.org/10.1016/j.wocn.2007.01.002.

Pouplier, M., & Beňuš, Š. (2011). On the phonetic status of syllabic consonants: Evidence from Slovak. *Laboratory Phonology, 2*(2). https://doi.org/10.1515/labphon.2011.009.

Pouplier, M., & Goldstein, L. (2010). Intention in articulation: Articulatory timing in alternating consonant sequences and its implications for models of speech production. *Language and Cognitive Processes, 25*(5). https://doi.org/10.1080/01690960903395380.

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. *Proceedings of Speech Prosody 2002*, 115–120. http://cogprints.org/2273/.

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

Rathcke, T., Falk, S., & Bella, S. D. (2021). Music to your ears: Sentence sonority and listener background modulate the "speech-to-song illusion". *Music Perception, 38*(5), 499–508. https://doi.org/10.1525/MP.2021.38.5.499.

Rathcke, T., & Lin, C.-Y. (2021). Towards a comprehensive account of rhythm processing issues in developmental dyslexia. *Brain Sciences, 11*(10), 1303. https://doi.org/10.3390/brainsci11101303.

Rathcke, T. V., Lin, C.-Y., Falk, S., & Dalla Bella, S. (2019). When language hits the beat: Synchronising movement to simple tonal and verbal stimuli. In *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1505–1509).

Rathcke, T. V., Lin, C.-Y., Falk, S., & Dalla Bella, S. (2021). Tapping into linguistic rhythm. *Laboratory Phonology Journal of the Association for Laboratory Phonology, 12*(1), 11. https://doi.org/10.5334/labphon.248.

Rathcke, T. V., & Smith, R. H. (2015). Speech timing and linguistic rhythm: On the acoustic bases of rhythm typologies. *The Journal of the Acoustical Society of America, 137*(5), 2834–2845. https://doi.org/10.1121/1.4919322.

Reddy, A. A., Chennupati, N., & Yegnanarayana, B. (2013). Syllable nuclei detection using perceptually significant features. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (pp. 963–967).

Repp, B. H. (2004). On the nature of phase attraction in sensorimotor synchronization with interleaved auditory sequences. *Human Movement Science, 23*(3–4 SPE. ISS.), 389–413. https://doi.org/10.1016/j.humov.2004.08.014.

Repp, B. H. (2005). Sensorimotor synchronization: A review of the tapping literature. *Psychonomic Bulletin and Review, 12*(6), 969–992. https://doi.org/10.3758/BF03206433.

Repp, B. H. (2006). Rate Limits of Sensorimotor Synchronization. *Advances in Cognitive Psychology, 2*(2–3), 163–181. https://doi.org/10.2478/v10053-008-0053-9.

Repp, B. H. (2010). Sensorimotor synchronization and perception of timing: Effects of music training and task experience. *Human Movement Science, 29*(2), 200–213. https://doi.org/10.1016/j.humov.2009.08.002.

Repp, B. H., & Su, Y.-H. (2013). Sensorimotor synchronization: A review of recent research (2006–2012). *Psychonomic Bulletin and Review, 20*(3), 403–452. https://doi.org/10.3758/s13423-012-0371-2.

Ridouane, R. (2002). Words without Vowels: Phonetic and Phonological Evidence from Tashlhiyt Berber 1. *ZAS Papers in Linguistics, 28*, 93–110. https://doi.org/10.21248/ZASPIL.28.2002.161.

Riesberg, S., Kalbertodt, J., Baumann, S., & Himmelmann, N. P. (2020). Using Rapid Prosody Transcription to probe little-known prosodic systems: The case of Papuan Malay. *Laboratory Phonology : Journal of the Association for Laboratory Phonology, 2020*, 11. https://doi.org/10.5334/labphon.192.

Ross, J. M., Iversen, J. R., & Balasubramaniam, R. (2016). Motor simulation theories of musical beat perception. *Neurocase, 22*(6), 558–565. https://doi.org/10.1080/13554794.2016.1242756.

Saltzman, E., Nam, H., Krivokapic, J., & Goldstein, L. (2008). A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *Proceedings of the 4th International Conference on Speech Prosody, SP 2008* (pp. 175–184).

Scott, S. K. (1998). The point of P-centres. *Psychological Research, 61*(1), 4–11. https://doi.org/10.1007/PL00008162.

Scott, D. R., Isard, S. D., & de Boysson-Bardies, B. (1985). Perceptual isochrony in English and in French. *Journal of Phonetics, 13*(2), 155–162.

Selkirk, E. O. (1982). The syllable. In H. van der Hulst & N. Smith (Eds.), *The Structure of phonological representations (part II)*. Dordrecht: Foris.

Selkirk, E. O. (1984). Phonology and Syntax: The relation between sound and structure. In *Current studies in linguistics series* (pp. 476).

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C. C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). TOBI: A Standard for Labeling English Prosody. *Second International Conference on Spoken Language Processing*.

Sonderegger, M., Stuart-Smith, J., Knowles, T., Macdonald, R., & Rathcke, T. (2020). Structured heterogeneity in Scottish stops over the twentieth century. *Language, 96* (1), 94–125. https://doi.org/10.1353/lan.2020.0003.

Sowiński, J., & Dalla Bella, S. (2013). Poor synchronization to the beat may result from deficient auditory-motor mapping. *Neuropsychologia, 51*(10), 1952–1963. https://doi.org/10.1016/j.neuropsychologia.2013.06.027.

Stupacher, J., Hove, M. J., Novembre, G., Schütz-Bosbach, S., & Keller, P. E. (2013). Musical groove modulates motor cortex excitability: A TMS investigation. *Brain and Cognition, 82*(2), 127–136. https://doi.org/10.1016/j.bandc.2013.03.003.

Šturm, P., & Volín, J. (2016). P-centres in natural disyllabic Czech words in a large-scale speech-metronome synchronization experiment. *Journal of Phonetics, 55*, 38–52. https://doi.org/10.1016/j.wocn.2015.11.003.

Su, Y. H., & Pöppel, E. (2012). Body movement enhances the extraction of temporal structures in auditory sequences. *Psychological Research, 76*(3), 373–382. https://doi.org/10.1007/s00426-011-0346-3.

Suni, A., Šimko, J., Aalto, D., & Vainio, M. (2017). Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech and Language, 45*, 123–136. https://doi.org/10.1016/j.csl.2016.11.001.

Tilsen, S. (2009). Multitimescale dynamical interactions between speech rhythm and gesture. *Cognitive Science, 33*(5), 839–879. https://doi.org/10.1111/j.1551-6709.2009.01037.x.

Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America, 134*(1), 628–639. https://doi.org/10.1121/1.4807565.

Thomassen, J. M. (1982). Melodic accent: Experiments and a tentative model. *The Journal of the Acoustical Society of America, 71*(6), 1596–1605. https://doi.org/10.1121/1.387814.

Toft, Z. (2002). phonetics and phonology of some syllabic consonants in southern British English. *ZAS Papers in Linguistics, 28*, 111–144. https://doi.org/10.21248/zaspil.28.2002.162.

Turk, A., Nakai, S., & Sugahara, M. (2012). Acoustic segment durations in prosodic research: A practical guide. In *Methods in Empirical Prosody Research, January 2006* (pp. 1–28). https://doi.org/10.1515/9783110914641.1.

Villing, R. C., Repp, B. H., Ward, T. E., & Timoney, J. M. (2011). Measuring perceptual centers using the phase correction response. *Attention, Perception, and Psychophysics, 73*(5), 1614–1629. https://doi.org/10.3758/s13414-011-0110-1.

Wagner, P., Cwiek, A., & Samlowski, B. (2019). Exploiting the speech-gesture link to capture fine-grained prominence impressions and listening strategies. *Journal of Phonetics, 76*. https://doi.org/10.1016/J.WOCN.2019.07.001 100911.

White, L., Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language, 66*(4), 665–679. https://doi.org/10.1016/j.jml.2011.12.010.

Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: Auditory-motor interactions in music perception and production. *Nature Reviews Neuroscience, 8*(7), 547–558. https://doi.org/10.1038/nrn2152.

Zhang, L., Xie, S., Li, Y., Shu, H., & Zhang, Y. (2020). Perception of musical melody and rhythm as influenced by native language experience. *J. Acoust. Soc. Am., 147*, EL385–EL390. https://doi.org/10.1121/10.0001179.

Zhang, L., Shu, H., Zhou, F., Wang, X., & Li, P. (2010). Common and distinct neural substrates for the perception of speech rhythm and intonation. *Human Brain Mapping, 31*(7), 1106–1116. https://doi.org/10.1002/hbm.20922.

Zhang, Y., & Glass, J. R. (2009). Speech rhythm guided syllable nuclei detection. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (pp. 3797–3800). https://doi.org/10.1109/ICASSP.2009.4960454.