

Testing an acoustic model of the P-center in English and Japanese

Tamara V. Rathcke,^{1,a)} Eline A. Smit,¹ Chia-Yuan Lin,² and Haruo Kubozono³

¹Department of Linguistics, University of Konstanz, Konstanz, Baden-Württemberg 78464, Germany

²Department of Psychology, University of Huddersfield, Huddersfield, Yorkshire HD1 3DH, United Kingdom

³Research Department, National Institute for Japanese Language and Linguistics, Tachikawa, Tokyo 190-8561, Japan

ABSTRACT:

The notion of the “perceptual center” or the “P-center” has been put forward to account for the repeated finding that acoustic and perceived syllable onsets do not necessarily coincide, at least in the perception of simple monosyllables or disyllables. The magnitude of the discrepancy between acoustics and perception—the location of the P-center in the speech signal—has proven difficult to estimate, though acoustic models of the effect do exist. The present study asks if the P-center effect can be documented in natural connected speech of English and Japanese and examines if an acoustic model that defines the P-center as the moment of the fastest energy change in a syllabic amplitude envelope adequately reflects the P-center in the two languages. A sensorimotor synchronization paradigm was deployed to address the research questions. The results provide evidence for the existence of the P-center effect in speech of both languages while the acoustic P-center model is found to be less applicable to Japanese. Sensorimotor synchronization patterns further suggest that the P-center may reflect perceptual anticipation of a vowel onset.

© 2024 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1121/10.0025777>

(Received 20 November 2023; revised 27 March 2024; accepted 8 April 2024; published online 19 April 2024)

[Editor: Sven Mattys]

Pages: 2698–2706

I. INTRODUCTION

The perceptual center (or the P-center) is defined as the subjective moment of occurrence and refers to the perceptual onset of an acoustic event.^{1,2} P-centers play an important role in synchronized speech behaviors, such as chanting or choral reading,^{3,4} by offering an acoustic landmark for temporal inter-speaker coordination of speech onsets.⁵ The idea of the P-center highlights a discrepancy that exists between the acoustic signal and its perception when it comes to the temporal structure of speech and other complex sounds.^{6,7} It has been suggested that the P-center reflects the rhythmic beat of spoken syllables^{2,8,12} similar to the beat in music^{13,15} and therefore constitutes a foundational unit of rhythm in speech. Since the beat is often defined as a regular pulse that structures time in complex sounds such as music,¹⁶ the P-center has sometimes been interpreted as evidence that temporal regularity may be perceptual rather than acoustic in nature.^{1,17}

The P-center effect was discovered and has been extensively studied with a series of isolated monosyllabic (rarely disyllabic) words^{1,2,9,12,18} which limits the scope of conclusions that can be made with regard to the relationships between acoustics and its perception. Early work acknowledged that the P-center may be “subject to phonological, semantic, or syntactic influences” (Ref. 1, p. 405) though to date, little research attention has been paid to the P-center effect in connected speech. To fill this gap, the present study investigated the P-center effect

in sentences of varying length and complexity by employing a rhythmic synchronization paradigm.¹⁹ The paradigm follows on from the early work on the P-center⁸ and reflects recent advances made in the cross-disciplinary study of rhythm by means of sensorimotor synchronization.^{20,21} In contrast to a previous implementation of this paradigm that required synchronization with one designated syllable in a series of spoken sentence repetitions,⁸ the present study generally asked participants to synchronize with the beat of a sentence played back to them on a loop. Previous research has established that the extended version of the paradigm appears intuitive to participants and produces replicable patterns.^{19,22} The main advantage of the new paradigm is its ability to be used with a great variety of natural, unmanipulated, complex sentences.

Existing investigations of the P-center effect focus primarily on simple, mostly monosyllabic speech materials and have adopted a wide range of methodologies, including perceptual adjustments for isochrony,^{2,11,23} speech production in time with a metronome,^{18,24,25} syllable repetition at self-paced, equidistant intervals,¹² as well as synchronizing finger taps with the syllable beat.⁸ Regardless of the deployed methodology, previous work has converged on one main finding that perceived onsets of spoken syllables tend to lag behind their acoustic onsets, though the exact location of the P-center has proven difficult to determine.¹⁵ It is generally assumed to be shaped by the properties of the syllable structure, yet no conclusive evidence exists. For example, some studies show that the P-center is influenced by onset consonants but not codas²³ while others provide evidence for an

^{a)}Email: tamara.rathcke@uni-konstanz.de

effect of both onsets and codas.²⁶ Some studies argue that only consonants but not vowels play a role in influencing the location of the P-center²⁷ while others observe a measurable impact of vowel quality on the P-center location.²⁵ The only widely accepted consensus of the research field suggests that the P-center approximates vowel onsets.^{12,23,28}

Despite the challenges in locating the P-center in the acoustic signal, the perceptual effect itself has been repeatedly attested in a variety of rhythmically distinct as well as tonal languages^{28,29} and is considered a cross-linguistic universal.¹² Given that the P-center effect persists despite substantial cross-linguistic differences in syllable phonology and prosodic structure, several existing models of the P-center plausibly assume that its location is best determined by the properties of the acoustic signal.^{7,18,30,31} Accordingly, the P-center effect arises as a consequence of the perceptual system sampling amplitude envelopes and responding particularly sensitively to salient, acoustically defined points at onsets of speech units (such as syllables). Exact acoustic definitions of salience differ across models,¹⁵ though many of them identify specifically the midpoints of amplitude rise-times as the most salient points of syllable envelopes.³² The most recent model of the P-center revisits this definition and argues that the model performs best if amplitude modulations of fricative onsets are removed from the calculation of the P-center.¹⁸ Instead, it is best represented as the moment of the fastest energy change that occurs at the consonant-vowel transition within a syllable. The energy change relates specifically to the amplitude rise of the vowel onset while amplitude changes due to high-energy consonants (such as fricatives) do not contribute to P-center location (Ref. 18, p. 42).

The model was developed for Czech and successfully tested with a large number of phonotactically varied, disyllabic words representative of Czech syllable complexity.¹⁸ The algorithm accompanying the model smoothes the raw amplitude envelope by calculating energy averages across 40-ms windows and applying a 44-sample shift and the 6th-order moving-average filter. Subsequently, it calculates energy differences between two neighbouring samples while disregarding samples with zero crossing rates higher than 4000 (i.e., fricatives) and smoothing the difference values via a moving-average filter of order 10 (see Fig. 1). Here, we apply the P-center model to two typologically, rhythmically, and phonologically distinct languages, English and Japanese. (British) English is a Germanic language that permits syllables of relatively high complexity in onset and coda position and includes lexical stress in its phonological representations.³³ (Tokyo) Japanese is a Japonic language with more restricted syllable phonotactics accompanied by a moraic quantity system and a lexically specified pitch accent.^{34,35}

One existing study has previously compared the P-center in these two languages.¹² Four speakers of each language had the task of producing pairs of monosyllables with different onsets, repeating them 10 times at a steady, comfortable pace, and maintaining isochronous intervals

between the repetitions. All monosyllables were phonologically licit nonce words of both languages, *a-ba*, *ma-ba*, and *pa-sa*. Analyses of durations measured between successive syllable vs vowel onsets indicated that speakers of both languages tended to produce more regular timings between vowel rather than syllable onsets. The results were interpreted as supporting “the idea that the P-center may be linked with the articulatory onset of the vowel” (Ref. 12, p. 373), though the exact location of the P-center could not be determined, given the methodology of the study. Further methodological restrictions to four speakers per language and five monosyllables with a simple consonant-vowel structure pose additional limitations to the generalizability of the P-center findings comparing English and Japanese.

The goals of the present study are twofold. First, it seeks to establish if the P-center effect exists in natural connected speech of two rhythmically distinct languages, English and Japanese, since previous work focused exclusively on isolated monosyllables.¹² To this end, we expect participants of both languages to display sensorimotor synchronization patterns consistently deviating from (i.e., lagging behind) syllable onsets. Second, the study aims to establish if the acoustic model of the P-center that captures the moment of the fastest energy change within a syllable as the core rhythmic unit¹⁸ can adequately capture the P-center effect in the two languages, typologically and rhythmically unrelated to the language for which the algorithm had been originally developed.¹⁸ To this end, we expect all participants’ synchronization to temporally coincide with the algorithm-derived P-center locations. If, however, the acoustic model does not adequately represent the P-center in either of these languages, we expect to find more consistent synchronization patterns with vowel onsets as the closest representatives of the P-center location.^{12,23,28}

II. METHODS

A. Participants

Thirty-six native British English speakers were recruited and tested at the University of Kent at Canterbury, England (22 female, M_{age} 25.5 and SD_{age} 5.92). Forty native Tokyo Japanese speakers were recruited and tested at the National Institute for Japanese Language and Linguistics and at the International Christian University in Tokyo, Japan as well as at the University of Huddersfield, England (27 female, M_{age} 21.8 and SD_{age} 4.21).

The participants reported variable levels of foreign language proficiency. Fourteen English participants had basic knowledge of Spanish, French, Italian, Japanese, German, Russian, or Armenian. Five Japanese participants were fluent in English, Korean, or Spanish while thirty participants reported basic knowledge of English or Korean. Three Japanese participants identified themselves as bilingual. We quantified the reported language profiles as 0 (no foreign or second language experience), 1 (basic knowledge), 2 (fluent), and 3 (multilingual), and checked if this experience influenced participants’ synchronization performance in the

present study,³⁶ but did not find any effects. The participants further reported their levels of previous music training that were quantified as an index following the procedure from previous research.^{18,19} The index was an aggregated score reflecting years of musical training, current regular music practice (0 or 1), number of musical instruments (including voice) and the age at which participants started playing (2 for early onset below 10, 1 for later onset between 10 and 20 years, and 0 for late onset above 20 years). No professional musicians or dancers took part in the experiment. Participants of both languages scored between 0 (no musical training) and 32 (extensive musical training), with no significant group-level differences between them. All participants were informed about the purpose of the experiment and signed a consent form prior to participation. The study received ethical approval from the Ethics Board of the University of Kent.

B. Materials

Twenty sentences per language were tested in the study. English sentences were selected from an existing database that had been created for a previous study.¹⁹ Japanese sentences were developed to mirror English sentences in length and complexity, by systematically matching the number of moras to the number of syllables in English materials. For example, the English sentence consisting of nine syllables “There will be a south-easterly wind.” was matched with the Japanese sentence consisting of nine moras “Onagadori-ga tonda.” (“A long-tailed bird flew.”) Sentence length ranged from minimally 4 to maximally 11 units (for the full list of sentences, see Ref. 37). Apart from being more naturalistic, the materials used in the present study comprised a greater diversity of syllable structures than previously investigated.¹² All sentences were recorded in a sound-attenuated booth at the University of Kent, with a sampling rate of 44 100 Hz, by native female speakers in their 20s who spoke Southern British English or Tokyo Japanese. The speakers were instructed to read the sentences in their comfortable speaking voice, without slowing down or hyper-articulating (for sound examples, see Ref. 37).

Vowel onsets and syllable boundaries of the English materials were manually annotated by a trained phonetician (the first author). The annotations of the Japanese materials were completed by a native Japanese phonetician and subsequently checked and corrected by the first author whenever necessary, to ensure a consistent application of the annotation criteria. Vowels were defined with reference to the presence of voicing, formant structure, and high intensity (excluding pre- or post-aspiration). If a vowel was preceded by a sonorant, an auditory criterion of the intended vowel quality was additionally applied. An epenthetic glottal stop was treated as an onset consonant (see “incident” in Fig. 1). Epenthetic glottal stops were frequent in onsetless English syllables. The materials comprised of 150 syllables in total, with 12 of them (8%) being onsetless and 6 (4%) containing an epenthetic glottal stop. In contrast, Japanese materials

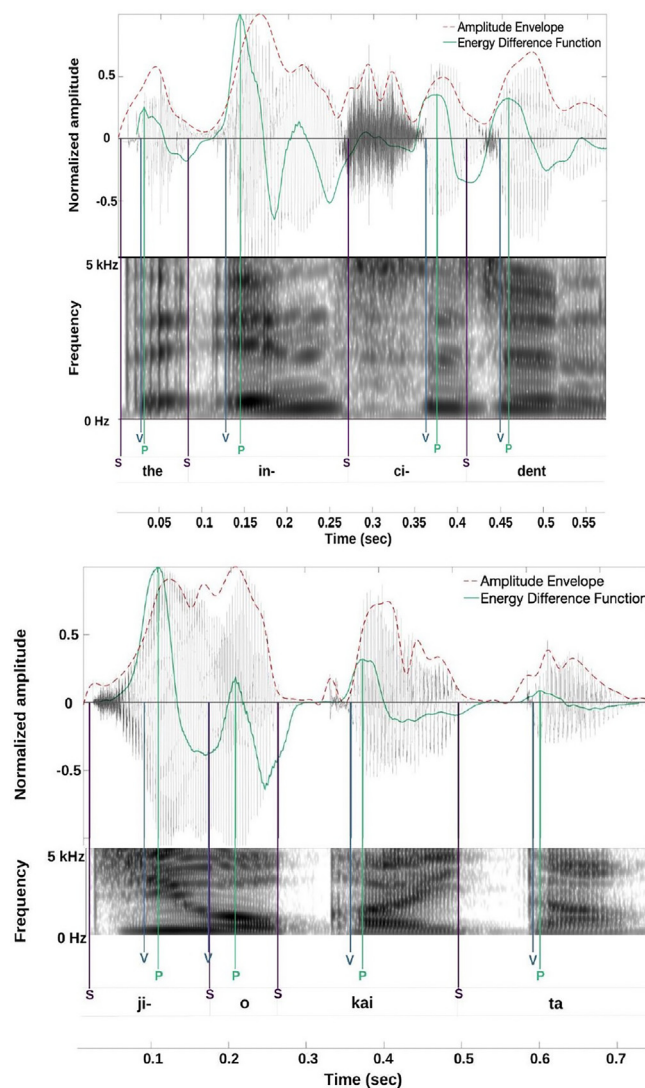


FIG. 1. (Color online) Waveforms, amplitude envelopes, energy difference functions, spectrograms, and example annotations of experimental materials comparing time points of manually identified syllable (S) and vowel (V) onsets alongside algorithm-derived P-center locations (P) in a quadrisyllabic English phrase taken from the test sentence “The incident occurred last Friday night” (top panel) and a quadrisyllabic Japanese test sentence “Ji-o kai ta” (“I wrote a character,” bottom panel).

included a higher number of onsetless syllables (18 out of 125, or 14.4%), and none of them featured an epenthetic consonant (e.g., Fig. 1), though epenthetic consonants may occur in Japanese onsetless syllables.³⁵ Example annotations are available from the OSF page of the project (see Ref. 37).

The P-center of each syllable was determined by an existing algorithm based on an acoustic model of the amplitude modulation within a syllable.¹⁸ The algorithm first created a raw energy contour of each sentence, by squaring raw amplitudes and using a 40-ms window and a 1-ms shift. A smoothed energy contour was then derived by applying a 6th-order moving average filter to the raw contour (cf. dotted lines in Fig. 1). Subsequently, energy dynamics of the smoothed contour were captured as an energy difference function. It was calculated by comparing the energy values

of two neighbouring samples ($E_t - E_{(t-1)}$), disregarding all samples with zero crossing rates higher than 4000 (i.e., fricatives), and smoothing the neighbouring difference values via a moving-average filter of order 10 (cf. solid lines in Fig. 1). The P-center corresponds to the moment of the fastest energy change within a syllable, represented by a local maximum of the energy difference function (cf. P in Fig. 1).

Figure 1 displays examples of the manual annotations of syllable and vowel onsets as well as the algorithm-derived P-center locations for (a) an English phrase and (b) a Japanese sentence, each containing four syllables of variable structure (including onsetless).

C. Procedure

Prior to the main experimental task, participants filled in an online questionnaire that collected demographic information on their linguistic and demographic background and screened for potential disorders of language, speech, or movement. They further completed a series of simple sensorimotor synchronization tasks with a metronome that ensured all participants had typically developing synchronization profiles and were able to perform the experimental task.³⁸

In the experimental task, participants were asked to listen to repetitions of a spoken sentence and to start synchronizing with what they perceived to be the beat of the sentence as soon as they had the feeling of the beat. They were instructed to tap in synchrony with the beat, using the finger of their dominant hand on a drumming pad placed in front of them. The instructions further required participants to maintain their synchronization until the sentence repetitions ended. Three practice trials were provided at the beginning of the experiment, including an opportunity to clarify questions with the experimenter and to adjust the volume of the playback to an individually comfortable level. Twenty sentences in participants' native language were presented in a randomized order. Each trial consisted of 15 repetitions of a sentence, separated by a 400 ms pause.¹⁹

The sentences were played back through good-quality headphones (Sennheiser HD 380 Pro). The timing of finger taps was collected on a Roland HPD-20 Handsonic Percussion Pad using a Dell Latitude 7390 laptop and CAKEWALK MIDI software (by BandLab).³⁹

D. Data pre-processing and statistical analyses

The timing of finger taps was extracted using the MATLAB MIDI toolbox⁴⁰ and corrected by subtracting the delay of the MIDI device (in this case, 5 ms). The timing of each tap was then mapped onto the three types of landmarks annotated or derived for each sentence (P-centers, syllable and vowel onsets). We first analysed the time course of taps produced throughout the series of sentence repetitions, starting from the first tap. For this, we applied a ± 120 ms window^{19, 21} and assigned every tap to each derived or annotated landmark located within the pre-defined window. The procedure derived signed asynchronies (in ms) with the three landmarks, resulting in negative values if a tap occurred before a

landmark and in positive values if a tap lagged behind a landmark.

The time-series of taps produced during each sentence loop were analyzed using the generalized additive mixed models (GAMMs), examining the relationship between signed asynchronies and the three landmarks. These models were estimated by the maximal likelihood method and included a random smoothing factor for each participant and sentence. The time-series factor was the serial order of taps recorded from the beginning until the end of each individual synchronization trial, with the first tap labeled as 1 (regardless of when a participant started synchronizing during sentence repetitions). The performance of English and Japanese participants was analyzed separately, to establish the time course of the potential P-center effect in the two languages and to answer to the first research question of the present study.

In these analyses, the effect is attested if taps consistently follow syllable onsets, i.e., when there is a typical discrepancy between perceptual and acoustic events.^{1,2,8} Given that the task requires some practice and may cause a period of variability prior to an individually consistent and reliable pattern of synchronization,¹⁹ it is primarily the timing of taps produced toward the end of a sentence loop that is of special interest here. The acoustic model of the P-center¹⁸ can be considered as an adequate representation of the P-center effect if the taps are timed with close reference to the derived P-landmark.

Apart from the time-series analyses, we examined which time-points within each sentence were consistently most likely to attract a tap throughout repetitions.¹⁹ For these likelihood-based analyses, a Gaussian kernel density estimation procedure (using ggplot2 in R⁴¹ and a bandwidth adjustment of $\frac{1}{8}$) was then applied. This procedure allows for a smooth distribution of all taps to be obtained for each participant and sentence. Individual densities were obtained from the synchronization data and aggregated across all participants (see Fig. 2). Local density peaks reflect points of high synchronization likelihood, characterizing rhythmic profiles of individual participants as well as sentences and helping to deal with the trial-by-trial variability during synchronization.^{19,42}

Local maxima in the individual density distributions were estimated using the *findpeaks* function from the R-package PRACMA.⁴³ The procedure applies a 40%-threshold of each sentence's maximum peak value and identifies all peaks above this threshold as the most likely points anchoring sensorimotor synchronization with the sentence. After having identified the temporal locations of density peaks, we calculated signed and absolute asynchronies with each landmark of interest (the P-center, vowel, and syllable onset) and applied the same ± 120 ms window for these asynchrony calculations as for the time-series based analyses. Both asynchronies measure the temporal distance between a landmark and a tapping peak (in milliseconds). While signed asynchronies indicate if tapping peaks tended to precede or follow landmarks and can be interpreted as an

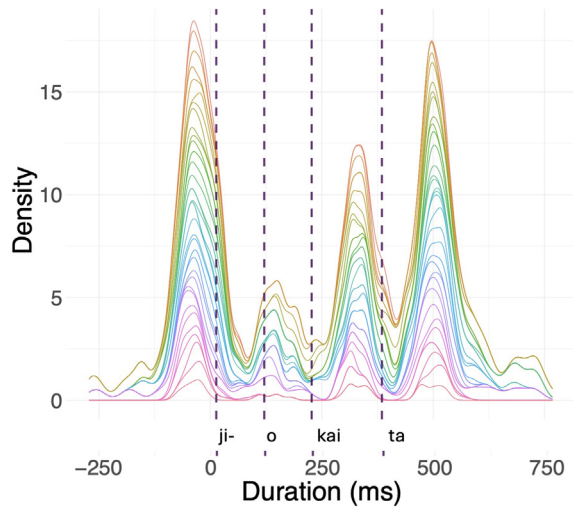


FIG. 2. (Color online) Example of a density function reflecting Japanese group performance with the test sentence “Ji-o kai ta” (“I wrote a character,” consisting of 4 syllables or 5 morae). Each coloured line represents tapping density of one participant. Sentence duration is plotted along the x axis, with 0 ms representing the onset of the sentence. Dashed vertical lines indicate syllable onsets.

index of temporal tracking vs anticipation, absolute asynchronies (i.e., the mode of intervals between tapping peaks and local landmarks) are a general index of synchronization error.²² Prior to statistical analyses, absolute asynchronies were log-transformed to reduce skewness.⁴⁴

Signed and absolute asynchronies derived from the density distributions were analyzed using linear mixed-effects models (LMEMs) and fit in R library lme4 environment.⁴⁵ The settings included the “optimx” optimizer and the base-R “nlminb” control option with the aim to resolve model convergence issues while keeping the random effects structure maximal.^{46,47} These models were fit to the English and Japanese data independently and tested for a main effect of landmark (the P-center, vowel, and syllable onset). All models included random terms for participant and sentence.

The model fit was estimated using the maximum likelihood test.⁴⁵

These analyses aimed primarily at answering the second research question of the present study. Accordingly, the acoustic model of the P-center¹⁸ can be considered as equally representative of the two languages under investigation if taps were timed (i.e., showed smallest absolute asynchronies) with the P-landmarks in both languages. Analyses of signed asynchronies can further help to confirm the conclusion based on the GAMMs analyses: the P-center effect is attested if taps were very likely to follow (rather than precede) syllable onsets.^{1,2,8}

III. RESULTS

A. Time-series based analyses

Figure 3 visualizes the results of the GAMMs analyses. It displays changes in signed asynchrony over the time course of sentence repetitions. As can be seen, English participants started the task by tapping slightly ahead of syllable onsets and quickly changed to lagging behind this landmark after a few synchronization cycles, thus showing a clear P-center effect in the second half of their synchronization. GAMMs identified that neither the intercept (β 0.33, SE 0.48, t 0.69, p 0.49) nor the smooth terms (edf 1.02, Ref.df 1.04, F 2.01, p 0.16) differed significantly between the acoustically derived P-center locations and the manually defined vowel onsets. In contrast, there was a significant difference between the P-center locations and the syllable onsets in both the intercepts (β 18.35, SE 0.49, t 37.42, $p < 0.0001$) and the smooth terms (edf 3.99, Ref.df 4.91, F 11.86, $p < 0.0001$).

Figure 3 further indicates that Japanese participants started tapping exactly at syllable onsets and shifted away from this landmark after just a few taps, thus showing a relatively early onset of the P-center effect. For this group, GAMMs identified that the intercept (β 9.18, SE 0.46,

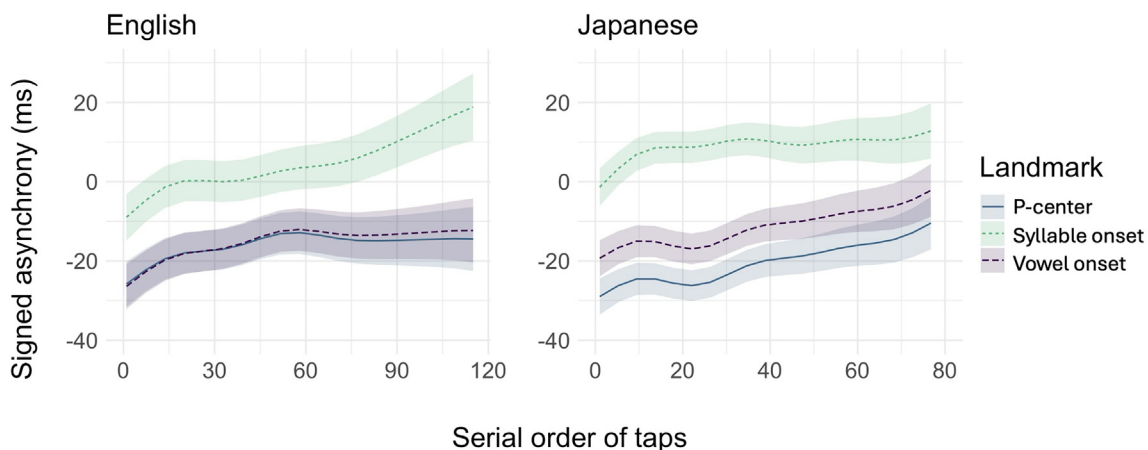


FIG. 3. (Color online) Results of GAMMs fit to the series of taps produced during repeated sentence synchronization by English (left panel) and Japanese (right panel) participants. The x axis reflects the order of taps recorded from the beginning of each experimental trial (i.e., 0 serial order). The y axis displays signed asynchronies (i.e., temporal distances between the timing of a landmark and a nearby tap) and is scaled around 0 ms, representing the time point of a landmark in the acoustic signal. Negative/positive values of a fitted curve indicate that taps preceded/followed a landmark.

$t = 20.16, p < 0.0001$) but not the smooth terms (edf = 1.00, Ref.df = 1.00, $F = 0.62, p = 0.43$) differed significantly between the acoustically derived P-center locations and the manually identified vowel onsets. Moreover, the P-center locations significantly differed from the syllable onsets in both the intercept ($\beta = 31.13, SE = 0.46, t = 67.43, p < 0.0001$) and the smooth term (edf = 4.58, Ref.df = 5.55, $F = 10.94, p < 0.0001$).

As in previous studies, we observed some individual variability in the onset of synchronization.¹⁹ Some participants started to synchronize early in the loop while others waited until the last few repetition cycles before starting to tap along. This variability in individual performance led to a decreasing amount of tapping data (and less reliable tapping trajectories) at higher serial orders. The plot in Fig. 3 is therefore capped at the first 120 taps for English participants and the first 80 taps for Japanese participants. Given that English vs Japanese materials were matched in the number of syllables vs moras, the syllable count of Japanese materials was lower, resulting in shorter tapping sequences as compared to English participants' data.

B. Likelihood-based analyses

Figure 4 plots estimated means and standard errors of signed and absolute asynchronies that were derived from tapping probabilities and calculated as a lag between a local tapping peak and the neighbouring P-center, vowel, or syllable onset. The interaction of landmark and language significantly improved model fit of both signed asynchrony ($F = 4.52, p = 0.011$) and absolute asynchrony ($F = 6.25, p = 0.0019$) models. To test the differences relevant to our research aims, we conducted pairwise comparisons corrected using the Tukey method as implemented in the emmeans package⁴⁸ in R.⁴⁹ Planned cross-linguistic comparisons indicated that both English ($\beta = 10.95, SE = 1.40$) and Japanese participants ($\beta = 7.09, SE = 1.49$) lagged behind syllable onsets during synchronization, though to a different extent ($\beta = 3.86, SE = 1.72, z = 3.86, p = 0.025$). Signed asynchronies measured with reference to the algorithm-derived P-center also showed substantial differences

between English and Japanese ($\beta = 9.90, SE = 1.72, z = 9.90, p < 0.0001$), with the P-center location being more representative of the synchronization performance found in English ($\beta = 2.40, SE = 1.39$) rather than Japanese ($\beta = 12.30, SE = 1.50$) participants.

These analyses can be compared to the raw tapping data shown in Fig. 5, demonstrating a more substantial overlap of asynchrony values measured in English vs Japanese participants' data.

Tables I and II provide a summary of within-language models, focusing on the planned landmark comparisons that best captured participants' performance. For English, syllable onsets were further away from tapping peaks than both P-centers ($z = 7.06, p < 0.0001$) and vowel onsets ($z = 6.99, p < 0.0001$). In contrast, absolute asynchronies measured with vowel onsets and P-centers did not differ ($z = 0.10, p = 0.99$). Signed asynchronies were also comparable for these two landmarks ($z = 1.12, p = 0.50$). For Japanese, we similarly found that absolute asynchronies were significantly higher for syllable onsets compared to both P-centers ($z = 4.69, p < 0.0001$) and vowel onsets ($z = 9.26, p < 0.0001$). Comparing vowel onsets and P-centers in Japanese, we found that vowel onsets were located closer to the peaks of tapping distributions than P-centers ($z = 4.52, p < 0.0001$), though the difference between the two landmarks was not significant for signed asynchronies ($z = 1.60, p = 0.25$).

IV. DISCUSSION

This research examined the P-center and its acoustic model in two rhythmically and typologically distinct languages, English and Japanese. The first goal of the present study was to establish if the P-center effect occurred in natural connected speech that was more complex than hitherto tested speech materials consisting of monosyllabic^{1,2,9,12} or (rarely) disyllabic¹⁸ words. To this end, we deployed a finger tapping paradigm^{20,21} and examined rhythmic synchronization with repeated sentences presented as short loops. The paradigm was developed in our previous work¹⁹ as an extension of early studies of the P-center that defined it as

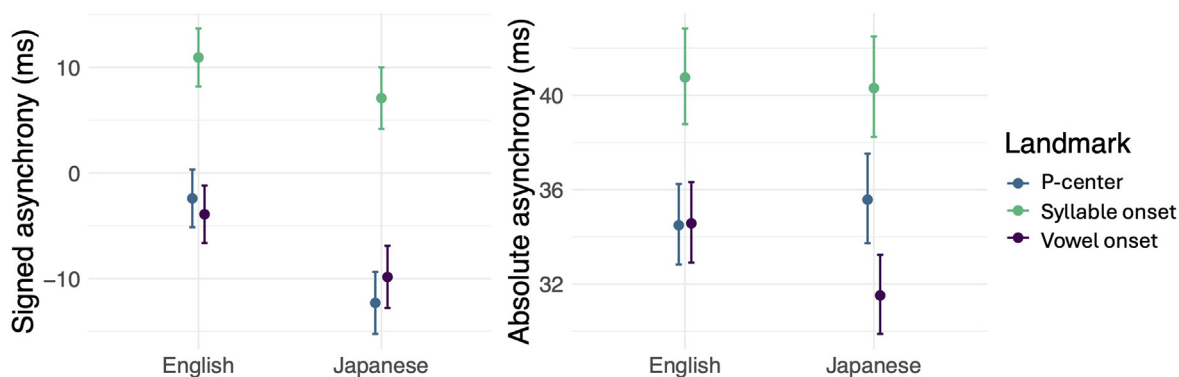


FIG. 4. (Color online) Estimated means and standard errors of signed asynchronies (left, reflecting temporal anticipation vs tracking) and absolute asynchronies (right, reflecting synchronization error) derived from tapping distributions for the three landmarks (P-centers, syllable and vowel onsets). Absolute asynchronies were log-transformed prior to statistical modelling and back-transformed to the original scale for the plot.

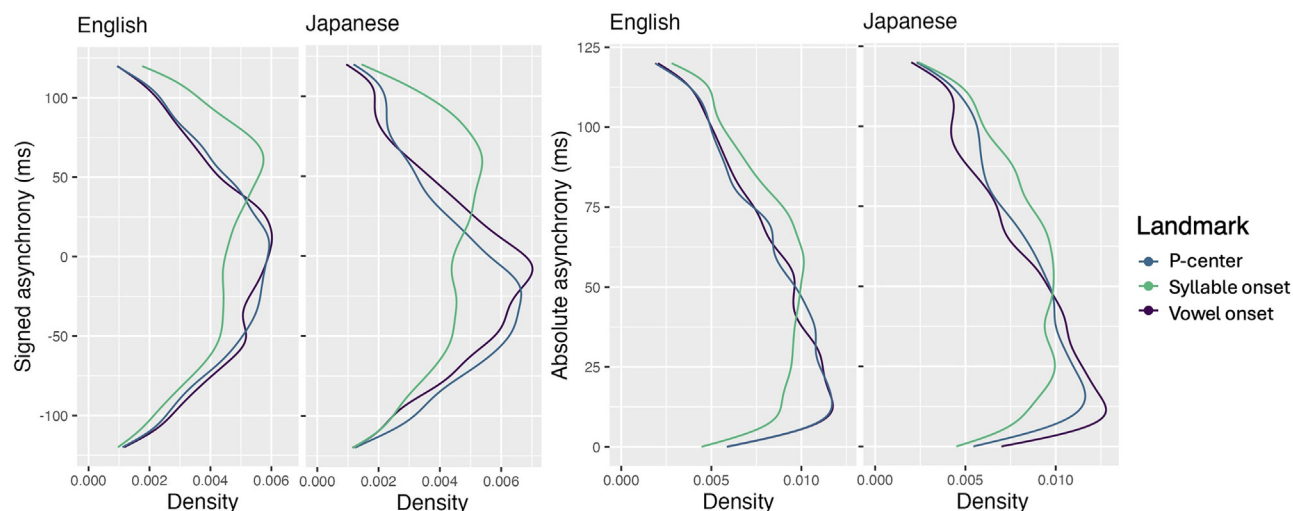


FIG. 5. (Color online) Density functions of raw tapping data plotting signed asynchronies (left panels) and absolute asynchronies (right panels), comparing the distribution across the three landmarks and the two languages.

the syllable beat.⁸ We investigated time-series as well as likelihood-based time-points of finger taps produced by English and Japanese participants who had completed the task of synchronizing with the beat of sentences spoken in their native language.

Time-series analyses were first applied to the timing of all finger taps produced during sentence loops, showing that both participant groups displayed the P-center effect by producing taps that consistently lagged behind syllable onsets.^{1,2,8} The likelihood-based analyses of synchronization time-points confirmed that taps had a strong tendency to occur after syllable onsets in both languages. This finding demonstrates that the P-center effect exists in connected speech of high levels of complexity. That is, the effect is not a mere artifact of methods that deploy repetitions of monosyllabic or disyllabic words interspersed with silent pauses. It further corroborates the previous conclusion that the P-center is a necessary component of temporally sensitive perception of any complex auditory stimulus.¹⁵

The second goal of the study was to establish if the recent acoustic model of the P-center¹⁸ would be applicable across a variety of languages. The model is based on an analysis of amplitude modulations of acoustic signal envelopes and defines the P-center as the moment of the fastest

energy change within a syllable.¹⁸ The rationale of the model is implemented in an algorithm that had been originally developed and tested with Czech, a Slavonic language that is typologically and rhythmically different from the languages of the present study. Using the algorithm, we derived P-center locations in twenty English and Japanese sentences of varying length and syllable structure. All analyses of the present study converged on the key finding that English participants' synchronization was equally well captured by algorithm-generated P-centers and manually annotated vowel onsets.^{19,22} In contrast, Japanese participants' synchronization was less well reflected in the algorithm-derived P-center locations, suggesting that the algorithm did not provide an adequate representation of the P-center in connected Japanese speech.

This discrepancy may arise from general cross-linguistic differences in syllable phonologies and consonant inventories of the two languages. Since many acoustic models of the P-center (including the model tested in the present study) sample amplitude envelopes specifically at syllable onsets,^{7,18,30,31} it is not unlikely that their P-center representations will be more applicable to languages rich in obstruents and complex onset clusters (like Czech and English) but less applicable to languages with fewer obstruents and more

TABLE I. Planned comparisons of estimated marginal means for English tapping data (likelihood-based analyses).

ENG	β	SE	z	p
Absolute asynchrony				
P-center - vowel onset	0.00	0.02	0.10	0.99
P-center - syllable onset	0.17	0.02	7.06	<0.0001
vowel onset - syllable onset	0.16	0.02	6.99	<0.0001
Signed asynchrony				
P-center - vowel onset	1.50	1.34	1.12	0.50
P-center - syllable onset	13.35	1.36	9.84	<0.0001
vowel onset - syllable onset	14.85	1.35	10.99	<0.0001

TABLE II. Planned comparisons of estimated marginal means for Japanese tapping data (likelihood-based analyses).

JAP	β	SE	z	p
Absolute asynchrony				
P-center—Vowel onset	0.12	0.03	4.52	<0.0001
P-center—Syllable onset	0.12	0.03	4.69	<0.0001
vowel onset—Syllable onset	0.25	0.03	9.26	<0.0001
Signed asynchrony				
P-center—Vowel onset	2.46	1.54	1.60	0.25
P-center—Syllable onset	19.39	1.53	12.69	<0.0001
vowel onset—Syllable onset	16.93	1.53	11.10	<0.0001

restrictive syllable phonotactics (like Japanese).^{34,35} Even though Japanese phonology includes several obstruents, its syllable structure licences onsets to maximally two consonants with the second position being occupied exclusively by the high-sonority approximant /j/. This property of the Japanese phonological system may lead to comparatively short periods of low-sonority in the acoustic signal. Low-sonority onsets (due to the presence of obstruents and their clusters) induce characteristic changes to amplitude envelopes⁵⁰ that can be easily conceptualized as modulations of the rate of amplitude change¹⁸ while high-sonority onsets or onsetless syllables (possibly accompanied by vowel hiatus) are less likely to leave a measurable imprint on the amplitude modulation at vowel onsets. Notably, half of phonologically onsetless syllables from our English set of sentences were produced with an epenthetic glottal stop, thus enhancing amplitude modulations at onsets of those syllables (see example in Fig. 1). Such realizations were absent in Japanese materials (see example in Fig. 1). Given these cross-linguistic differences in syllable phonology and their substantial acoustic consequences, it may be not surprising that there exists “no widely accepted and generally applicable acoustic P-center model” (Ref. 15, p. 1615). Future approaches to an automatic calculation of the P-center location may benefit from moving toward more linguistically informed accounts and from potentially considering auditory effects that are at play during the perception of acoustic signals.^{30,31}

Finally, present findings indicate that rhythmic synchronization in the two languages of the study may be best captured as being timed with vowel onsets, corroborating a previous conclusion (made on the basis of less naturalistic speech materials) that in many languages, the P-center approximates a vowel onset within a syllable unit.^{12,23,28} Here, we also observed a tendency for participants’ taps to precede synchronization targets by a few milliseconds on average—a phenomenon commonly found in synchronization studies and referred to as negative mean asynchrony.^{20,51} Reasons for negative mean asynchrony are still poorly understood, though it has sometimes been suggested to be indicative of temporal anticipation that is central to the sensorimotor synchronization ability.⁵¹ Accordingly, finger taps precede rather than follow synchronization targets to ensure temporal alignment between two involved systems that differ in the speed of transmission, somatosensory feedback from finger taps and auditory feedback from rhythmic events. As seen in the time-series analyses of the tapping data, participants’ synchronization tended to stabilize toward the end of a trial in close proximity, yet ahead of, a vowel onset. This suggests that the P-center effect may in fact reflect an anticipated onset of a vowel as a primarily relevant constituent of syllable, the smallest structuring unit of prosodic hierarchy.²²

V. CONCLUSION

To conclude, the results of the present study provided evidence for the P-center effect in natural connected speech

of English and Japanese, extending previous findings obtained in the two languages with simple, meaningless, monosyllabic sequences.¹² The applicability of a recent acoustic model of the P-center¹⁸ was, however, shown to be restricted to English as its potential to account for the Japanese data of the present study was limited. We suggest that the P-center reflects perceptual anticipation of a vowel onset.

ACKNOWLEDGMENTS

This research was supported by a research grant from the Leverhulme Trust (Grant No. RPG-2017-306) and a JSPS visiting fellowship to T.V.R. The authors thank Ryuichi Taki (NINJAL) for his help with the annotations of Japanese materials.

AUTHOR DECLARATIONS

Conflict of Interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethics Approval

The study obtained ethical approval from the Ethics boards of all involved institutions and the experiment was performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all individual participants included in the study.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

- ¹J. Morton, S. Marcus, and C. Frankish, “Perceptual centers (P-centers),” *Psychol. Rev.* **83**(5), 405–408 (1976).
- ²S. K. Scott, “The point of P-centres,” *Psychol. Res.* **61**(1), 4–11 (1998).
- ³F. Cummins, “Rhythm as entrainment: The case of synchronous speech,” *J. Phon.* **37**(1), 16–28 (2009).
- ⁴K. Cerda-Oñate, G. T. Vega, and M. Ordín, “Speech rhythm convergence in a dyadic reading task,” *Speech Commun.* **131**, 1–12 (2021).
- ⁵A. D. MacIntyre, C. Q. Cai, and S. K. Scott, “Pushing the envelope: Evaluating speech rhythm with different envelope extraction techniques,” *J. Acoust. Soc. Am.* **151**(3), 2002–2026 (2022).
- ⁶J. London, K. Nymoen, M. T. Langerød, M. R. Thompson, D. L. Code, and A. Danielsen, “A comparison of methods for investigating the perceptual center of musical sounds,” *Atten. Percept. Psychophys.* **81**(6), 2088–2101 (2019).
- ⁷J. Vos and R. Rasch, “The perceptual onset of musical tones,” *Percept. Psychophys.* **29**(4), 323–335 (1981).
- ⁸G. D. Allen, “The location of rhythmic stress beats in English: An experimental study I,” *Lang. Speech* **15**(1), 72–100 (1972).
- ⁹R. Cumming, A. Wilson, V. Leong, L. J. Colling, and U. Goswami, “Awareness of rhythm patterns in speech and music in children with specific language impairments,” *Front. Hum. Neurosci.* **9**, 672 (2015).
- ¹⁰C. A. Harsin and K. P. Green, “Perceptual centers as an index of speech rhythm,” *J. Acoust. Soc. Am.* **96**(5), 3350 (1994).
- ¹¹C. A. Harsin, “Perceptual-center modeling is affected by including acoustic rate-of-change modulations,” *Percept. Psychophys.* **59**(2), 243–251 (1997).

- ¹²C. Hoequist, Jr., "Syllable duration in stress-, syllable- and mora-timed languages," *Phonetica* **40**(3), 203–237 (1983).
- ¹³A. Danielsen, K. Nymoen, M. T. Langerød, E. Jacobsen, M. Johansson, and J. London, "Sounds familiar(?): Expertise with specific musical genres modulates timing perception and micro-level synchronization to auditory stimuli," *Atten. Percept. Psychophys.* **84**(2), 599–615 (2022).
- ¹⁴R. C. Villing, "Hearing the moment: Measures and models of the perceptual Centre," Ph.D. thesis, National University of Ireland, Galway, Ireland, 2010.
- ¹⁵R. C. Villing, B. H. Repp, T. E. Ward, and J. M. Timoney, "Measuring perceptual centers using the phase correction response," *Atten. Percept. Psychophys.* **73**(5), 1614–1629 (2011).
- ¹⁶P. E. Savage, S. Brown, E. Sakai, and T. E. Currie, "Statistical universals reveal the structures and functions of human music," *Proc. Natl. Acad. Sci. U.S.A.* **112**(29), 8987–8992 (2015).
- ¹⁷I. Lehiste, "Isochrony reconsidered," *J. Phon.* **5**(3), 253–263 (1977).
- ¹⁸P. Šturm and J. Volín, "P-centres in natural disyllabic Czech words in a large-scale speech-metronome synchronization experiment," *J. Phon.* **55**, 38–52 (2016).
- ¹⁹T. Rathcke, C.-Y. Lin, S. Falk, and S. D. Bella, "Tapping into linguistic rhythm," *Lab. Phon.: J. Assoc. Lab. Phonol.* **12**(1), 11 (2021).
- ²⁰B. H. Repp, "Rate limits in sensorimotor synchronization with auditory and visual sequences: The synchronization threshold and the benefits and costs of interval subdivision," *J. Motor Behav.* **35**(4), 355–370 (2003).
- ²¹B. H. Repp and Y.-H. Su, "Sensorimotor synchronization: A review of recent research (2006–2012)," *Psychon. Bull. Rev.* **20**(3), 403–452 (2013).
- ²²T. Rathcke and C.-Y. Lin, "An acoustic study of rhythmic synchronization with natural English speech," *J. Phon.* **100**, 101263 (2023).
- ²³S. M. Marcus, "Acoustic determinants of perceptual center (P-center) location," *Percept. Psychophys.* **30**(3), 247–256 (1981).
- ²⁴B. Tuller and C. A. Fowler, "Some articulatory correlates of perceptual isochrony," *Percept. Psychophys.* **27**, 277–283 (1980).
- ²⁵R. A. Fox and I. Lehiste, "The effect of vowel quality variations on stress-beat location," *J. Phon.* **15**(1), 1–13 (1987).
- ²⁶A. M. Cooper, D. H. Whalen, and C. A. Fowler, "The syllable's rhyme affects its P-center as a unit," *J. Phon.* **16**(2), 231–241 (1988).
- ²⁷A. Cooper, D. Whalen, and C. Fowler, "P-centers are unaffected by phonetic categorization," *Percept. Psychophys.* **39**, 187–196 (1986).
- ²⁸K. Franich, "Tonal and morphophonological effects on the location of perceptual centers (P-centers): Evidence from a Bantu language," *J. Phon.* **67**, 21–33 (2018).
- ²⁹Y.-J. Lin and K. de Jong, "The perceptual center in Mandarin Chinese syllables," *J. Phon.* **99**, 101245 (2023).
- ³⁰P. Howell, "Prediction of P-center location from the distribution of energy in the amplitude envelope: II," *Percept. Psychophys.* **43**(1), 99–99 (1988).
- ³¹S. K. Scott and P. Howell, "Perceptual centers in speech: An acoustic analysis," *J. Acoust. Soc. Am.* **92**, 2443 (1992).
- ³²F. Cummins and R. Port, "Rhythmic constraints on stress timing in English," *J. Phon.* **26**(2), 145–171 (1998).
- ³³D. Kahn, *Syllable-Based Generalizations in English Phonology* (Routledge, Abingdon, UK, 2015).
- ³⁴T. J. Vance, *An Introduction to Japanese Phonology* (State University of New York Press, Albany NY, 1987).
- ³⁵H. Kubozono, *Handbook of Japanese Phonetics and Phonology* (De Gruyter Mouton, Berlin, 2015).
- ³⁶P. Lidji, C. Palmer, I. Peretz, and M. Morningstar, "Listeners feel the beat: Entrainment to English and French speech rhythms," *Psychon. Bull. Rev.* **18**, 1035–1041 (2011).
- ³⁷T. Rathcke, E. Smit, and C.-Y. Lin, "Testing an acoustic model of the P-center in English and Japanese," <https://osf.io/whzqb> (2024) (Last viewed April 15, 2024).
- ³⁸S. Dalla Bella, N. Farrugia, C. E. Benoit, V. Biegel, L. Verga, E. Harding, and S. A. Kotz, "BAASTA: Battery for the assessment of auditory sensorimotor and timing abilities," *Behav. Res.* **49**, 1128–1145 (2017).
- ³⁹Cakewalk Inc., "Cakewalk by bandlab," (2019), <https://www.cakewalk.com/> (Last viewed December 1, 2019).
- ⁴⁰T. Eerola and P. Toivainen, *MIDI Toolbox: MATLAB Tools for Music Research* (University of Jyväskylä, Jyväskylä, Finland, 2004).
- ⁴¹H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer, Cham, 2016).
- ⁴²D. Dotov, D. Bosnyak, and L. J. Trainor, "Collective music listening: Movement energy is enhanced by groove and visual social cues," *Quart. J. Exp. Psychol.* **74**, 1037–1053 (2021).
- ⁴³H. Borchers, "PRACMA: Practical numerical math functions," <https://cran.r-project.org/package=pracma> (2018) (Last viewed November 10, 2023).
- ⁴⁴R. H. Baayen, *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*, 1st ed. (Cambridge University Press, Cambridge, 2008).
- ⁴⁵D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.* **67**(1), 1–48 (2015).
- ⁴⁶J. C. Nash and R. Varadhan, "Unifying optimization algorithms to aid software system users: Optimx for R," *J. Stat. Softw.* **43**(9), 1–14 (2011).
- ⁴⁷J. C. Nash, "On best practice optimization methods in R," *J. Stat. Softw.* **60**(2), 1–14 (2014).
- ⁴⁸R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means* (2022), r package version 1.7.3, <https://CRAN.R-project.org/package=emmeans> (Last viewed November 10, 2023).
- ⁴⁹R. C. Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org/> (2022) (Last viewed October 31, 2023).
- ⁵⁰L. Varnet, M. C. Ortiz-Barajas, R. Guevara Erra, J. Gervain, and C. Lorenzi, "A cross-linguistic study of speech modulation spectra," *J. Acoust. Soc. Am.* **142**(4), 1976–1989 (2017).
- ⁵¹G. Aschersleben, "Temporal control of movements in sensorimotor synchronization," *Brain Cogn.* **48**(1), 66–79 (2002).