Pushing the envelope: Evaluating speech rhythm with different envelope extraction techniques

Alexis Deighton MacIntyre, Ceci Qing Cai and Sophie K. Scott

Citation: The Journal of the Acoustical Society of America **151**, 2002 (2022); doi: 10.1121/10.0009844
View online: https://doi.org/10.1121/10.0009844
View Table of Contents: https://asa.scitation.org/toc/jas/151/3
Published by the Acoustical Society of America

JASA ARTICLE

Pushing the envelope: Evaluating speech rhythm with different envelope extraction techniques

Alexis Deighton MacIntyre,[a] Ceci Qing Cai, and Sophie K. Scott
Institute of Cognitive Neuroscience, University College London, London, WC1N 3AZ, United Kingdom

ABSTRACT:
The amplitude of the speech signal varies over time, and the speech envelope is an attempt to characterise this variation in the form of an acoustic feature. Although tacitly assumed, the similarity between the speech envelope-derived time series and that of phonetic objects (e.g., vowels) remains empirically unestablished. The current paper, therefore, evaluates several speech envelope extraction techniques, such as the Hilbert transform, by comparing different acoustic landmarks (e.g., peaks in the speech envelope) with manual phonetic annotation in a naturalistic and diverse dataset. Joint speech tasks are also introduced to determine which acoustic landmarks are most closely coordinated when voices are aligned. Finally, the acoustic landmarks are evaluated as predictors for the temporal characterisation of speaking style using classification tasks. The landmark that performed most closely to annotated vowel onsets was peaks in the first derivative of a human audition-informed envelope, consistent with converging evidence from neural and behavioural data. However, differences also emerged based on language and speaking style. Overall, the results show that both the choice of speech envelope extraction technique and the form of speech under study affect how sensitive an engineered feature is at capturing aspects of speech rhythm, such as the timing of vowels. © *2022 Acoustical Society of America*. https://doi.org/10.1121/10.0009844

(Received 13 July 2021; revised 5 March 2022; accepted 5 March 2022; published online 23 March 2022)
[Editor: James F. Lynch] Pages: 2002–2026

I. INTRODUCTION

Derived from an ancient Greek word meaning "to flow," rhythm can be literally defined as the "manner of flowing" (Benveniste, 1971), but rhythm also means different things to different people. Among speech scientists, for example, there are probably as many definitions for rhythm as there are research groups. This presents a nontrivial challenge to interdisciplinary dialogue, in part, because differing definitions will naturally lead to differing methodologies. For instance, whereas the concept of rhythm may extend to any identifiable temporal pattern, other interpretations of rhythm mean an isochronous, or regularly timed, time series specifically. Within linguistics, this latter definition was associated with the controversial idea, known as the rhythm class hypothesis, that all languages can be grouped according to an isochronous or near-isochronous organising rhythmic unit, such as the syllable (Abercrombie, 1964). Theories of speech rhythm that centre the syllable require that this unit be well-defined, yet the formal characteristics of the syllable are subject to debate among phoneticians (Cummins, 2012b; Strauß and Schwartz, 2017; Zec, 2007), and determining precise syllabic boundaries can be especially difficult in spontaneous speech (Schachtenhaufen, 2010; Schuppler, 2017; Schuppler *et al.*, 2011). In any case, rhythm typologies have received little support from empirical studies (Arvaniti, 2009, 2012; Nolan and Jeon, 2014), and alternative metrics based on other linguistic constructs, such as ratios between the durations of consonant and vowel segments in speech (e.g., Grabe and Low, 2008; Ramus *et al.*, 1999), have also been shown to have poor predictive power (Arvaniti, 2012; Wiget *et al.*, 2010). Despite mixed evidence for the syllable as the fundamental unit in speech timing, let alone the ambiguity of the syllable itself, these concepts draw increasing interest within speech psychology and neuroscience. For example, recent advances in the analysis of the time-locked brain response to speech have enabled researchers to correlate between components of the acoustic stimulus and neural dynamics, leading to prominent theories of cortical speech tracking. These theories propose an oscillatory mechanism, meaning that some stimulus component needs to recur periodically for neural dynamics to track, align with, and even anticipate the speech time series (Ghitza, 2011, 2013; Giraud and Poeppel, 2012; Peelle and Davis, 2012). This component is often assumed to be the syllable, bringing us back to the rhythm class hypothesis and its implications concerning isochrony, the idea that speech is formed from regularly timed units (Cummins, 2012b). An important distinction between linguistic and neuropsychological studies of speech rhythm is that the method primarily favoured in neuroscience is to automatically extract acoustic features, usually some form of the speech envelope, for the purposes of correlation with signals recorded from a listener's brain. The theoretical justification for using acoustic features in speech perception experiments relies on the tacit assumption that these automatically generated components are an adequate stand in for phonetic

[a]Also at: MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge CB2 7EF, UK. Electronic mail: a.macintyre.17@ucl.ac.uk

analysis. In some papers, conclusions are drawn concerning the link between neural activity and the "syllabic rhythm" of speech stimuli, yet the methods sections exclusively describe acoustic feature extraction (e.g., Pefkou *et al.*, 2017; Vander Ghinst *et al.*, 2019). Other authors are more direct in stating that "vowels [correspond to] peaks in the speech envelope" (Ding and Simon, 2014). Although this substitution between linguistic concepts and acoustic features may be common practice, the empirical equivalence of phonetic tokens, such as syllables or vowels, and acoustically derived events, like peaks in the speech envelope, is not well established. To compound this uncertainty, the techniques used to generate acoustic features vary and cannot always be easily reproduced from the methods reported. It is, therefore, often difficult to gauge what the consequences of slightly differing algorithms or parameters may be. Given that experimental data within both linguistics and neuroscience are frequently measured at the time scale of tens of milliseconds, even small variations could lead to different scientific conclusions.

The common goals shared across linguistics, neuroscience, and other speech sciences should generate fruitful discussions. To encourage further dialogue and debate, it would be helpful to have a joint understanding of the basic aspects of terminology, for example, what is specifically captured by the "speech envelope," and how best to generate it, given the research goals at hand. Furthermore, it is important to establish how closely engineered features like the speech envelope can reflect the time series of phonetically defined concepts, such as vowels. Finally, differing automatic methods abound in the literature and the unique effects of these methods are largely unexplored. The aim of the current paper is, thus, to empirically establish how similar the information captured by acoustic landmarks, such as peaks in the speech envelope, is to that described via phonetic analysis. We also selected several different examples of speech envelope extraction techniques from the literature, including pipelines used in both linguistics (e.g., Tilsen and Arvaniti, 2013) and brain sciences (e.g., Oganian and Chang, 2019). By generating different versions of the speech envelope from the same annotated speech corpus, we attempt to quantify the extent to which methodological choices can affect the resultant time series. In sum, this project is motivated to characterise the similarities and discrepancies between phonetic annotation and engineered features, as well as evaluate the impact of specific techniques used to extract such engineered features. Ideally, standardising essential tools, like the speech envelope, will help researchers in one domain (e.g., neural speech perception) address the predictions of another (e.g., the basis of syllable-based timing), and vice versa. Although we cannot offer a definitive explanation of what speech rhythm is nor how it should be measured (Arvaniti, 2009; Kohler *et al.*, 2009), it is clear that disparate communities within speech rhythm research have each engaged with the syllable in some way or another. Whether or not these theories are ultimately proven or disproven will depend on the accurate and robust estimation of the syllabic time series. We, therefore, consider the precise determination of syllable timing as important to these questions but remain agnostic as to the legitimacy of their premise.

In the current paper, we combine phonetic speech rhythm transcription with different automatically extracted acoustic features that are frequently employed in the speech sciences. Using a purpose-bespoke corpus collected to emphasise the different contexts of speech rhythm, we ask what is commonly captured across these techniques. Moreover, we additionally include co-operative speaking conditions wherein two speakers read aloud in synchrony, allowing us to identify acoustic landmarks that are closely interpersonally coordinated but which may not have a direct analog in traditional transcription techniques. Together, these two tasks shed complementary light on the validity of automatically detected acoustic landmarks for speech timing research: on the one hand, we identify landmarks that most closely emulate a target of phonetic annotation, the vowel onset, which we have theoretical reasons to believe are behaviourally relevant to the percept of speech rhythm (Scott, 1998). On the other hand, the joint speech condition allows us to detect the acoustic anchors or targets that best facilitate interpersonal sensorimotor coordination during speech production. We identify the acoustic landmarks, as well as the specific parameters used to generate them, that best fulfil these criteria. As a preliminary step toward assessing their appropriateness for characterising aspects of speech rhythm, we apply this subset of selected acoustic landmarks in a proof of concept, a set of classification tasks that uses the time series of landmarks to predict the identities of different speaking styles (e.g., reading versus spontaneous speech). We perform these tasks using partitioned data from speakers unseen during the initial analyses, allowing us to provisionally establish the predictive utility of the parameters employed. This work, hence, contributes a groundwork for making objective choices in place of arbitrary or guesswork-based estimates when extracting algorithmically derived acoustic landmarks. The paper is structured as follows: we give an overview of the techniques for describing speech rhythm and its analysis across language, brain, and behavioural sciences, highlighting some methodological differences that impede comparisons across disciplinary boundaries. We then lay out our experimental approach, define the acoustic landmarks of interest, and state our procedure and analysis. We report our interim results, apply selected acoustic landmarks in a set of proof of concept speech classification tasks, and, finally, discuss our findings with particular emphasis toward its implications for studies in cortical speech tracking and rhythmic entrainment to speech more generally.

A. Background

1. Linguistic and behavioural analysis of speech timing

In music and movement sciences, research into rhythm has centred around the structure of regularly timed recurring

intervals, such as the beat by which dancers pace their steps. Although these behaviours can also encompass more nuanced forms of timing, a temporally regular structure underlies many musical and kinematic rhythms. For example, musical melodies tend to unfold against a referent musical pulse (i.e., the beat), and the temporal relationship between respiratory and gait cycles can often be expressed as a simple ratio, e.g., 2:1 (Bramble and Carrier, 1983). In the case of speech, however, such regularities can be harder to find; for example, there is no clear evidence for equally timed intervals between stressed syllables. For more exhaustive reviews of these efforts, readers are directed to Arvaniti (2009, 2012) and Cummins *et al*. (2009), but a rough consensus among linguists is that the syllabic time series or any variation thereof is too irregular to be considered as a recurring rhythmic unit (Meyer *et al*., 2020). Indeed, the considerable variability present in proposed rhythmic units prompted the "iconoclastic view" that speech itself may be "inherently antirhythmic" (Nolan and Jeon, 2014). Alternative methods have been proposed for capturing rhythmic variation between different languages, for example, by contrasting the durations of voiced versus unvoiced segments (Wiget *et al*., 2010). One method is the normalised pairwise variability index (nPVI; Grabe and Low, 2002), which is calculated by measuring the difference between successive pairs of vowel durations, divided by the mean of those two durations. Similar to simple interval-based speech rhythm metrics, the nPVI demonstrates weak predictive power for classification (Arvaniti, 2012; Nolan and Asu, 2009). A drawback to speech rhythm metrics is that they attempt to describe motor behaviour that is both produced and perceived on the time scale of the psychological present (i.e., <5 s; Fraisse, 1978) by using aggregate statistical measures, which by nature require data points to be temporally decontextualised and pooled to calculate. Any global inferences that are drawn concerning how speech rhythm operates without regard for its embodied and temporal situatedness may, therefore, be misleading. Moreover, these metrics convey information about durational timing, which may be neurally (Breska and Ivry, 2016; Teki *et al*., 2011; Teki *et al*., 2012) and behaviourally (Pope and Studenka, 2019; Tierney and Kraus, 2015) distinct from the form of timing more closely associated with rhythm and motor sequencing, which is event timing (see Leow and Grahn, 2014, for review). An intuitive example of the difference between durational and event timing can be borrowed from tennis, where you would use durational timing to measure the length of time required to perform a serve and event timing to describe the pattern of recurring shots between two players engaged in a rally. The latter form is anticipatory, meaning that the timing of previous events is informative of how future events will be timed.

One behaviourally informed approach to speech rhythm, which is also related to event timing, is the study of perceptual centres, known as *p*-centres. *p*-centres are typically defined as the precise moment at which an event is perceived to have occurred, although they are also relevant to motor production, for instance, as temporal articulatory targets (Marcus, 1981). The basic concept of *p*-centres is agnostic to domain or modality, but *p*-centres in speech have probably received the most scrutiny. Their location can be inferred experimentally, for example, by asking participants to align two repeating syllable stimuli so that they seem to be regularly spaced in time (e.g., Scott, 1998), or synchronise their own finger tapping to a syllable that repeats on loop (e.g., Vos *et al*., 1995). The acoustic correlates of *p*-centres are not decisively modelled (Villing, 2010); however, experimental findings indicate consistent *p*-centre placement in languages as diverse as Czech (Šturm and Volín, 2016), Bantu (Franich, 2018), Brasilian Portuguese (Barbosa *et al*., 2005), and Japanese (Hoequist, 1983). Evidence from these studies suggests that *p*-centres lie close to the vowel onset or within the transition between a syllable-initial consonant and the vowel (Hoequist, 1983; Patel *et al*., 1999; Scott, 1998; Šturm and Volín, 2016), although Cantonese speakers by contrast appear to place the *p*-centre at the syllable rather than vowel onset (Chow *et al*., 2015). A limitation to *p*-centre research is that most behavioural paradigms make use of simple repeating, isolated syllables, and, thus, the behavioural relevance for *p*-centres in connected, naturalistic speech is unknown. Recently, Rathcke *et al*. (2021) found that when asked to tap along to complete utterances, participants were most likely to target vowel onsets in comparison to several other linguistic and acoustic landmarks that the authors measured. Although the stimuli were cycled in this case and, therefore, still unrepresentative of communicative speech in everyday life, this work nonetheless demonstrates that the *p*-centre phenomenon is not limited to single syllables or words, and individuals are consistent in their responses during sensorimotor synchronisation to complex utterances (Rathcke *et al*., 2021). These experimental data are also corroborated by ecological findings from the Amazonian Bora language, which suggest that the timings of vowels in the spoken form shape temporal structure in its natural, drummed analog (Seifart *et al*., 2018).

2. Neuroscientific accounts of timing in speech perception

Thus far, we have touched on linguistic-phonetic and psychophysical-behavioural approaches to understanding speech rhythm. We, now, take the perspective of brain sciences, especially the role of syllable timing in neural speech processing. Despite the lack of consensus among linguists with regard to basic rhythm units in speech, some researchers in the cognitive neuroscience community have identified periodicity, usually that of the inter-syllable time series, as a putative mechanism underlying speech perception (Ding and Simon, 2014; Giraud and Poeppel, 2012; Gross *et al*., 2013; Peelle and Davis, 2012). To simplify, it is hypothesised that populations of neurons synchronise their activity at the level or some harmonic of the so-called "speech rate." Known as neural entrainment, this theory proposes that the

synchrony between a listener's brain and slow components of an incoming speech signal facilitate the anticipation, tracking, chunking, and decoding of speech in real time (Ding *et al.*, 2016; Gross *et al.*, 2013; Meyer and Gumbert, 2018; Meyer *et al.*, 2017; Molinaro and Lizarazu, 2018). Although neural entrainment as a mechanism underlying speech perception was predicted as early as the 1970s (Jones, 1976), there are some issues that have yet to be resolved (Cummins, 2012a,b). Importantly, the speech stimuli used in many perception experiments have employed synthetic syllables whose durational variabilities and timing do not reflect natural speech (Meyer *et al.*, 2020). Hence, the supposed irregularity present in everyday speech challenges oscillatory theories of speech perception to account for the flexibility of any interunit but especially intersyllable intervals (Doelling and Assaneo, 2021; Ghitza, 2013; Strauß and Schwartz, 2017).

Many experiments investigating the neural basis of speech perception use some form of the speech envelope (sometimes referred to as the amplitude envelope, temporal envelope, or intensity contour), a smoothed signal conveying the slow amplitude modulations within the speech wave form (Ding *et al.*, 2017). In the literature, this term has become a catchall for similar acoustic features extracted from speech recordings or used to synthesize stimuli, which can then be tested for correspondences to physiological signals that originate in the brain. These responses are typically measured by electroencephalography (EEG) or magnetoencephalography (MEG). The resultant correlations, phase-locking, and/or peaks in the frequency-domain that are common to both the stimuli and brain response may be interpreted as evidence for neural entrainment, whether as an oscillatory or time-locked stimulus response. Indeed, these measures are not necessarily informative of how entrainment is actually being driven; for instance, the exact definition of entrainment and whether or not entrainment necessarily entails oscillatory activity, or rather can be considered as evoked, is subject to debate (Doelling and Assaneo, 2021; Meyer *et al.*, 2020; Zoefel *et al.*, 2018). In any case, when researchers use speech envelopes to form conclusions about neural processes during speech perception, they typically engage with data and theory from linguistics, for example, by drawing connections between the average inter-event timing of syllables or stressed syllables with generic EEG/MEG frequency bins, such as the theta (∼4–8 Hz) or delta (∼1–3 Hz) bands (e.g., Ghitza, 2013; Giraud and Poeppel, 2012; but see Keitel *et al.*, 2018). As with the rhythm metrics identified in linguistics, these estimates often reflect aggregate statistics that are calculated from many observations collapsed across time. Nonetheless, neuroscientific experimental inferences that draw from linguistic theory are only as valid as the measures on which they rest, in this case, acoustic landmarks. What remains unclear is how closely phonetic annotation and automatic techniques resemble each other in practice and to what extent temporal nuance is captured using the latter.

3. Acoustic landmarks for speech rhythm analysis

To better generate and test predictions concerning how speech is processed in the brain and how we are able to coordinate our speech together in everyday conversation, linguistic, behavioural, and neuroscientific approaches to speech rhythm research should be reconciled, but this will require some synthesis of disparate theories and methods. In particular, it is necessary to establish the effect of the specific pipeline and parameters used to produce commonly used features, such as the speech envelope. The current paper is, thus, motivated to quantify how phonetic annotations compare with the automatically extracted features, which are increasingly favoured both in psychology and neuroscience, as well as across speech sciences more generally. There is currently little standardisation or consensus as to how speech envelopes should be extracted. Biesmans *et al.* (2017) compared among envelope extraction techniques in the context of auditory attention decoding, and report enhanced classification when listeners' brain responses were compared to envelopes calculated with "auditory-inspired modifications," such as gammatone filterbanks, in contrast to simple half-wave rectification (e.g., Dellwo *et al.*, 2015; Kolly and Dellwo, 2014) or by taking the analytic signal via Hilbert transform (e.g., Gervain and Geffen, 2019; Presacco *et al.*, 2016), two commonly used means to derive the envelope (Ding *et al.*, 2017). The findings by Biesmans *et al.* (2017) suggest that for the purposes of investigating the cortical tracking of speech, engineered features are not all built alike. Yet, different forms of phonetic annotation may also vary in their validity or relevance for speech rhythm research not withstanding complexities and unresolved debate concerning the form and boundaries of the syllable (Goldsmith, 2011; Zec, 2007). The question, therefore, arises as to what linguistic ground truth should acoustic landmarks be compared?

Given the behavioural evidence, we covered previously, *p*-centres present as an ideal candidate acoustic landmark by which the brain response can become paired with or entrained to the speech signal. In this respect, *p*-centres remain largely unexplored, although previous work has indicated the perceptual relevance of related features, deemed "acoustic edges" (Ding and Simon, 2014). More recently, evidence from electrocorticography (ECoG) demonstrates that activity within a defined region of the superior temporal gyrus (STG) may correspond to the *p*-centre time series (Oganian and Chang, 2019). Specifically, the authors found that the STG encodes the timing and magnitude of the speech amplitude envelope rate of change, such that steeper slopes (i.e., the most change in the least amount of time) elicit greater cortical responses (Oganian and Chang, 2019; Yi *et al.*, 2019). Notably, this result held across English, Spanish, and Mandarin stimuli, and the neural activity between English and Spanish monolingual participants to the same stimuli did not differ (Oganian and Chang, 2019). Following the phonetic analysis of their stimuli corpus, the authors showed that local maxima in the envelope rate of

change closely correspond to vowel onsets, particularly stressed vowel onsets in the case of English speech. In a subsequent preprint, using naturalistic stimuli at slowed and normal speeds, the authors observed that this same acoustic marker explained neural phase alignment to the speech signal (Kojima *et al.*, 2021). Taken together with the aforementioned behavioural accounts, these data potentially constitute convergent neural evidence for the relevance of *p*-centres to the analysis of speech rhythm. We, thus, focus on vowel onsets as the phonetic target in this study (Rathcke *et al.*, 2021; Scott, 1998).

Hence, the objective of the current paper is to characterise the similarities between the time series of acoustic landmarks, generated using engineered features, and the time series of vowel onsets, phonetically identified events. In doing so, we estimate the extent to which different envelope extraction techniques actually correspond to each other. Moreover, we can also gauge the impact of algorithmic parameters in landmark detection (e.g., minimum height for peak detection), as these specific steps are not often reported in the literature. Although previous studies have explored automated vowel onset annotation procedures (e.g., Adi *et al.*, 2016; Kumar *et al.*, 2017), most work is derived from relatively homogeneous corpuses consisting of read laboratory speech (e.g., TIMIT; Garofolo *et al.*, 1993) or single words or phonemes with little temporal variation (Schuppler, 2017). Moreover, even within read speech, the choice of text materials within languages has been shown to influence measures of rhythm over and beyond supposed linguistic differences (Wiget *et al.*, 2010). Here, we employ a multi-speaker, multilingual data set featuring three forms of naturalistic, connected speech that differ in rhythmic character. This corpus was custom collected to test the performance of different algorithms across diverse speech that was collected under comparable speaking and acoustic recording conditions.

In the spirit of *p*-centre research, additionally, we take an implicit approach to behavioural salience in speech rhythm by including *joint speech* in our experimental corpus. Joint speech entails two or more speakers joining their voices together at the same time in close synchrony, which is typically performed at a high level of precision without practice or training (Cummins, 2014, 2019). It is unclear exactly how this is achieved, but working under the assumption that more perceptually salient landmarks will be more closely coordinated than less perceptually salient landmarks, we can follow the data to identify which acoustic speech events are associated with the smallest asynchronies between speakers. The joint speech analysis allows for other acoustic landmarks to demonstrate potential utility in speech rhythm research, especially those that do not necessarily correspond to the selected *a priori* ground truth, vowel onsets (i.e., the approximate phonetic correlate of *p*-centres; Rathcke *et al.*, 2021). This aspect of the current study is motivated by the idea that linguistic constructs, such as vowels, may not represent a 1:1 correspondence to the acoustic features that support speech rhythm entrainment from a neurobiological perspective (Cummins, 2012b; Strauß and Schwartz, 2017), especially when we consider that humans begin life without knowledge of any written language system let alone what exactly constitutes a syllable (Räsänen *et al.*, 2018). To produce a wider range of acoustic landmarks for this more exploratory part of the study, we also include a set of complementary auditory features, consisting of gammatone cepstral coefficients.

In summary, we identify acoustic landmarks, generated by automatically extracting signal events from engineered features, that best estimate manually annotated vowel onsets. Furthermore, we describe landmarks on a corresponding time scale, which are closely coordinated between speakers who are attempting to read aloud together in synchrony. The predictive utility of these acoustic landmarks is explored using a proof of concept, a set of classification tasks comparing between the different forms of speech included in the experimental corpus. Our goal is to provide the speech sciences community with a quantitative comparison between different options for engineered features and ultimately help bridge the gap between the brain-based, behavioural, and linguistic approaches to understanding speech rhythm.

II. THE CURRENT STUDY

A. Overview

1. Corpus

The corpus used to extract and compare features is a balanced data set consisting of English and Mandarin speech, which has been theorised to differ in temporal organisation (e.g., Lin and Wang, 2007). Each language has seven speakers who were grouped into four dyad pairs per language for the purposes of joint speech (with one speaker per language performing in multiple dyads). The speakers each contributed matching solo and joint speech trials that included popular science articles adapted for length and ease of reading, formally structured poetry typical of its linguistic-cultural context, and spontaneously produced speech that was seeded by a semi-structured interview format. Although they differ on a number of continua, these speaking conditions were primarily selected to evoke diverse forms of temporal organisation. Part of the corpus (four of seven speakers per language) was manually annotated by acoustic syllable onset, vowel onset, and stressed vowel onset, resulting in >20 000 vowel/syllable and nearly 11 000 stressed vowel signal events recorded. Because *p*-centres are roughly localised to the vowel onset (Scott, 1998; Villing, 2010), and the prosodic role of stress in Mandarin remains unestablished (Duanmu, 2001; Lai *et al.*, 2010), we analyse only vowel onsets herein.

The speakers whose data were partially annotated ($n = 4$ per language) formed the development set, 80% of which was used to determine which acoustic landmarks most closely emulated vowel onsets with 20% held out to confirm the results. The remaining speakers ($n = 3$ per language) were set aside as a test set to apply the chosen

https://doi.org/10.1121/10.0009844

acoustic landmarks in a proof of concept speech analysis using machine learning classification. Our acoustic analysis can be summarised in three parts: (1) initial acoustic landmark identification and selection for estimation of manual vowel onset annotations, (2) selection of landmarks with smallest between-speaker asynchronies in joint speech trials, and (3) application of all selected landmarks in a classification task.

2. Acoustic feature extraction and landmark identification

After partitioning our corpus into development and testing sets, we first extracted, from the raw acoustic development data, candidate features in which to search for signal events that would be relevant to our research goals, namely, signal events that mimic humanly produced annotations and signal events that minimise the asynchronies between speakers during joint speech. The candidate features consisted of four different algorithms to calculate the speech amplitude envelope plus a set of gammatone cepstral coefficients (GTCCs), which were primarily included to produce a wider range of acoustic landmarks that could be tested in the joint speech analysis. Gammatone filters model the human auditory response, specifically, the spectral analysis performed by the cochlea; GTCCs can, therefore, be considered to be "biologically inspired" modifications of feature extraction techniques that decompose an input signal into the time-frequency domain (e.g., mel frequency cepstral coefficients; MFCCs), and which are highly popular in the audio processing world for their representation of complex signals, such as speech with substantially alleviated computational costs (Shao and Wang, 2008; Valero and Alias, 2012; Zhao and Wang, 2013). Cepstral coefficients remain widely unexplored in human behavioural research despite their ubiquitous application in technological systems, for instance, in voice recognition systems. Following feature extraction, we then used different event-finding algorithms to determine peaks and other acoustic landmarks, such as peaks in the first derivative, within those features. For each feature and landmark combination, we explored a variety of algorithmic parameters (e.g., the minimum temporal interval between peaks) that were narrowed down by an iterative, data-driven approach.

3. Acoustic landmark selection and applications

Based on how well the generated landmarks approximated the manual vowel onset annotations, we chose a subset of high-performing combinations of features and landmarks combinations ("acoustic landmarks") to carry forward. In a subset of joint speech trials only, we also extracted asynchronies in Euclidean distance between acoustic landmarks of the speakers (for manual vowel onset annotations as well as landmarks) using a mutual two-way closest match pairing algorithm. This measure permits us to compare among different types of discrete events in the speech signals between two joint speakers and thereby ascertain whether some acoustic landmarks are more closely coordinated than others. We, again, selected a subset of tightly synchronised acoustic landmarks from this joint speech analysis to take forward. The results from both steps were then confirmed in the 20% of development data that were held out.

After choosing acoustic landmarks based on approximation of manual vowel onset annotations and coordination between speakers in joint speech, we calculated various descriptive statistics, including, for instance, median and coefficient of quartile variation, from windowed inter-event interval data in the hitherto unseen test set speakers and applied these as predictors in a support-vector machine (SVM) classification task to discriminate between different types of speech rhythm, for example, solo versus joint reading. Following these three stages of analysis, the results are discussed in the context of speech rhythm, neural entrainment, and advancing the dialogue concerning phonetic and neurobiological theories of speech rhythm perception and production.

B. Methods

1. Speech recording

Participants were tested at the Institute of Cognitive Neuroscience. The acoustic speech signal was sampled at 44 100 Hz using SM58 cardioid dynamic microphones (Shure Inc., Niles, IL), positioned via a microphone stand in front of the speaker's mouth. In the case of joint speech, each participant was recorded using either the left or right channel. The first session was always the "solo speech" condition, and the second session was always the "joint speech" condition. As joint speech is likely to be more cognitively demanding than reading aloud alone, this order of trials was chosen to improve performance in the dyadic condition via a presumed practice effect. Stimuli texts consisted of adaptations of popular science articles matched for reading level, content, and tone; two poems characterised as typical of metred poetry according to the linguistic context (each read twice to match the duration of the articles); and a variety of prompts for spontaneous speech. The articles were edited to avoid potentially unfamiliar or phonetically ambiguous words or Chinese characters, and the order (articles; poems; spontaneous) was held constant. The text materials are available in Appendix D. Participants were verbally prompted when to begin speaking, and read from large-type, printed texts displayed on a stand at a comfortable reading distance, approximately 75 cm. Dyads were positioned side by side, reading from the same stand, separated by a distance of approximately 150 cm. We recorded during setup to ensure there was minimal bleed between the speakers' microphones, and re-positioned the speakers if necessary. The experimenter was present in the room but did not face the participants during recording. Extended instances of complete discoordination (i.e., one speaker drops out for longer than 1 s) were removed from the joint data set.

2. Preprocessing and annotations

A portion of the acoustic recordings were exported as *.wav files to be analysed in Praat (Boersma and Weenink, 2020). The speech data were annotated by A.D.M. and C.Q. C. with help from English- and Mandarin-speaking undergraduate volunteers. Vowel onsets were determined as the early emergence of strong formants in the broadband spectrogram. In the case of approximants and nasal consonants, the coincidence between the increase in intensity in the wave form alongside the appearance of higher formants in the spectrogram was chosen. The authors reviewed all annotations, and a consensus was reached where disagreement arose with A.D.M. performing a final inspection for consistency. The resulting manual vowel onset annotation time series, as well as the annotated and unannotated trials, were imported to MATLAB (The MathWorks, Natick, MA). Prior to feature extraction, silent periods (>500 ms) were truncated, and the acoustic data were windowed using a custom script that searched for natural break-points in silent periods (>100 ms), optimised for finding windows of 4 s in length but permissive within a range of 3–6 s. The mean duration was 4.01 s (standard deviation, SD = 0.57). This duration was determined to roughly balance the count of observations within windows with the overall sample size of windows. The data from four speakers per language were used for the acoustic landmark selection process (development). The development data were further partitioned with 20% held out to confirm the results. The remaining data from the three speakers from each language were used in evaluating the final choice of landmarks in the speech rhythm classification tasks (testing).

3. Acoustic feature extraction

The speech envelopes were calculated using four techniques, each of which has been previously published in the literature, from engineering to cognitive neuroscience to linguistics. We summarise the differing methods here.

(1) We employ taking the moving max of the absolute values of the signal using a 250 ms moving window. This method is most similar to simple signal rectification-based approaches and is proposed to avoid the attenuation seen when the Hilbert technique is applied to complex naturalistic sounds, such as speech or music (Caetano and Rodet, 2011; Jarne, 2018);
(2) we compute the magnitude of the Hilbert transform, which is used extensively across the speech entrainment literature (e.g., Assaneo *et al.*, 2019; Braiman *et al.*, 2018; O'Sullivan *et al.*, 2015). We took the absolute values from the output of the Hilbert function in MATLAB (The MathWorks, Natick, MA), but this algorithm is also implemented in Praat (He and Dellwo, 2016);
(3) we adopt the method described by Oganian and Chang (2019), adapted from Schotola (1984), which extracts the envelope from critical bands in the speech signal based on the Bark scale in an effort to emulate human audition (Zwicker and Terhardt, 1980; Zwicker *et al.*, 1979). The signal is first rectified within each filter bank and then averaged across all frequency bands; and
(4) we filter the speech signal using a fourth-order bandpass Butterworth filter at [400,4000] Hz, the estimated locus of vocalic energy (Tilsen and Arvaniti, 2013).

In each case, the resultant speech envelopes were smoothed using zero-phase low-pass filters with a 10 Hz cut-off and rescaled to a common range to facilitate comparison. Code to produce these speech envelopes is available for download.[1]

To widen the variety of acoustic landmarks for the joint speech analysis, we additionally extracted GTCC auditory features using the MATLAB function gtcc( ) (The MathWorks, Natick, MA). Initially, 13 coefficients were obtained (Revathi *et al.*, 2018). We then applied the feature selection to reduce this number, using principal component analysis to minimise computational costs and redundant information in the signals (Xie *et al.*, 2016). We found that, on average, the first three principal components explained 89.3% (SD 1.8%) of the total variance, and the first three GTCC features contributed the most to each of these components, hence, we retained GTCC 1–3. On visual inspection, we noticed that GTCC 1 also closely corresponded to the speech envelopes and, therefore, included it as an acoustic feature in the manual vowel estimation analysis. All of the extracted features had a sampling rate of about 660 Hz and were detrended and rescaled to [−1,1]. Noise in the signal floor was smoothed to avoid spurious peak detection using a custom script with a moving minimum mechanism. The pipeline used to produce the acoustic features is shown in Fig. 1(A). An example of the acoustic features used for vowel estimation (GTCC 1 and envelopes 1–4) alongside manually annotated vowel onsets is given in Fig. 2.

4. Signal event detection

We identified five different candidate signal events in each of the windowed features. In most of the cases, these events require specific parameters, such as the prominence of peaks. Here, we iterated over various parameter values for each discrete event, selecting those combinations of parameters that best solved the optimisation problem of matching vowel onsets in the current data set. Although we confirm the appropriateness and robustness of the final parameters by holding data out, the values we report will nonetheless be, to a certain extent, subject to the specifics of the current corpus, recording conditions, and our method of manual annotation. We enumerate the different signal events tested as follows:

(1) Lower crossing. Treating the acoustic feature as a bi-level signal consisting of a low and high phase, the moment of lower crossing refers to the instant that the positive-going signal crosses the lower state reference level, which was arbitrarily chosen as 10%. Detecting this transition from the low to high signal phase is

https://doi.org/10.1121/10.0009844

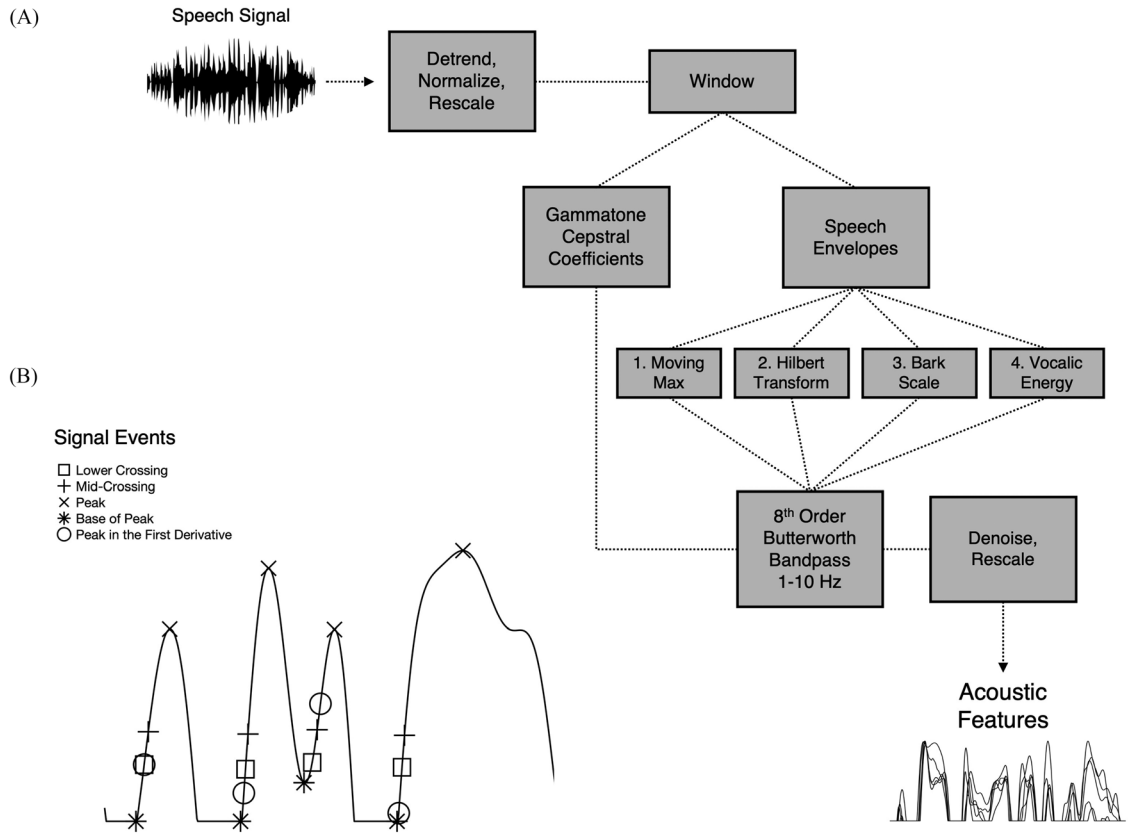

FIG. 1. (A) A block diagram of the process used to produce the acoustic features from the raw speech wave form and (B) example placements of the signal events are shown.

subject to a tolerance threshold, which was optimised on an iterative basis. We used the MATLAB (The MathWorks, Natick, MA) function risetime to estimate lower crossings.

(2) Mid-crossing. The mid-reference level crossing is calculated similarly to that in (1), except that the halfway or 50% point between the low and high signal phases rather than the crossing of the lower state (10%) is calculated.

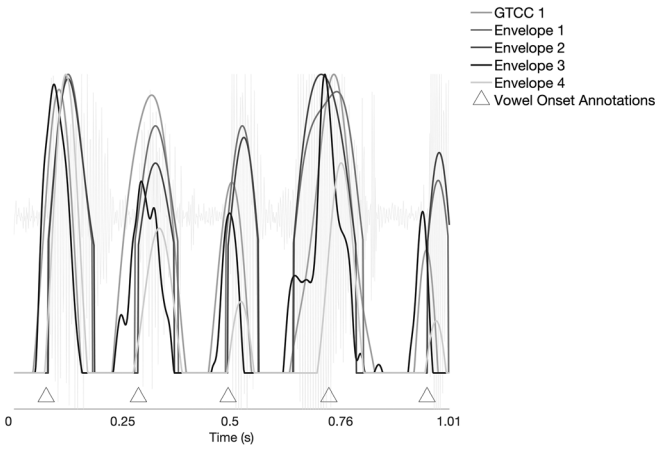

FIG. 2. The four automatically extracted acoustic features employed in our vowel onset estimation analysis, plotted against the raw speech wave form. The timing of the manual vowel onset annotations are represented by triangle markers.

(3) Peaks. To identify peaks in the feature, we used the islocalmax( ) function, iterating over topographic prominences ranging between [0,1] in increments of 0.1 and minimum inter-peak intervals ranging from [0,100] ms in increments of 20 ms.

(4) Base of peaks. The findchangepts( ) function, which locates abrupt changes in a signal, was applied in conjunction with islocalmax( ) to approximate the base of a positive-going slope leading to a peak.

(5) Peaks in the first derivative (rate of change) of the signal. This landmark is produced in the same way as that in (3), except that the input vector is the first derivative of the signal.

In total, for vowel onset estimation, we evaluated 5 features (4 speech envelopes and 1 GTCC) $\times$ 5 signal events, resulting in 25 different acoustic landmarks. For the joint speech analysis, there were 7 features (4 speech envelopes and 3 GTCC) $\times$ 5 signal events, resulting in 35 different acoustic landmarks. Figure 1(B) depicts example placements of the five different signal events.

5. Evaluation of acoustic landmarks

a. Vowel onset estimation. We first quantified the similarities between the linear acoustic features. If the results of the distinct speech envelope extraction methods differ only trivially, we should expect to see very high correlation

values between different speech envelopes generated from the same speech. To test this possibility, we took the acoustic features that had been extracted from each speech window and calculated the Pearson's r values between them. We then pooled the r values from across the individual windows and recorded the mean of these values as an indication of the overall linear correspondence between the acoustic features.

For each acoustic landmark, to determine which parameters best estimated manual vowel onset annotations, first, we compared the count of landmarks identified in each window with the ground truth, which was the corresponding count of manual vowel onset annotations. If the count of landmarks fell within a threshold of $\pm 10\%$ of ground truth, we awarded that window a score of one (otherwise, zero). The score was pooled first within speaking condition (articles, poems, or spontaneous speech) and then combined as a mean to produce a portion of matching counts. For each acoustic landmark (i.e., each unique combination of feature and signal event), we only took forward the highest scoring parameters (i.e., each unique combination of algorithmic parameters, such as inter-peak distance or peak prominence).

Following this first step, we then calculated the error as the Euclidean distance between the retained landmarks and annotations on a window by window basis via an in-house, mutual two-way closest matching script. This process produced three outcome variables: error values in milliseconds; the percentage of unpaired landmarks per total landmarks per window (i.e., false positives); and the percentage of unpaired annotations per total annotations per window (i.e., false negatives). Acoustic landmarks associated with mean unpaired percentages $>20\%$ were excluded outright.

We combined the outcome variables into a weighted score such that

$$\begin{aligned}\text{Vowel Estimation Score} &= 0.3 \times (1 - w_1) + 0.25 \times (1 - w_2) \\ &\quad + 0.45 \times (1 - w_3), \end{aligned} \tag{1}$$

where w_1 is the percentage of unpaired landmarks, w_2 is the percentage of unpaired annotations, and w_3 is the mean median error value. Weighting was determined with the aim to err on the side of fewer landmarks as higher counts of landmarks will likely produce smaller error values. The weighted score was calculated, first, within speaking condition (articles; poems; spontaneous) before being aggregated by acoustic landmark. Based on the mean weighted score, we identified three acoustic landmarks per language to carry forward in our applied classification task.

We also compared the distributional parameters of some of the selected inter-acoustic landmark intervals to those of the inter-annotation intervals. This was accomplished by calculating descriptive statistics from the inter-landmark and inter-annotation intervals for each speech window. These statistics included count, mean, mean absolute deviation, coefficient of variation, median, median absolute deviation, and coefficient of quartile variation. We then obtained r values by correlating between the statistics generated from the landmarks and those generated from annotations, allowing us to assess how closely the distributions of inter-acoustic landmarks follow those of annotations across the windowed speech.

b. Joint speech. Recording our speakers in pairings as they performed the reading tasks synchronously allowed us to ask whether some acoustic speech landmarks were more closely coordinated between participants (i.e., associated with smaller asynchronies and fewer unpaired speech events) than others, constituting an implicit, behavioural means of measuring the temporal salience of acoustic landmarks. To this end, we took the same acoustic landmarks that formed the first selection in Sec. II B 5 a (based on similar counts of landmarks versus corresponding manual annotations in windowed speech segments), but this time, we calculated the Euclidean distance in milliseconds between speakers (asynchrony) rather than between landmark and manual vowel onset annotations (error). For reference, we also calculated between-speaker asynchronies on the corresponding manual vowel onset annotation data. We selected the three most closely coordinated acoustic landmarks in each language, according to

$$\text{Joint Speech Score} = 0.6 \times (1 - w_1) + 0.4 \times (1 - w_2), \tag{2}$$

where w_1 is the mean unmatched landmark value across both speakers, and w_2 is the mean median asynchrony (ms) value between speakers.

In summary, a total of up to 12 unique acoustic landmarks (6 best vowel onset estimation; 6 most closely coordinated between speakers) could be selected based on the vowel onset estimation analysis and the joint speech analysis.

c. SVM classification of speech rhythm. As a proof of concept, we explored the application of the selected acoustic landmarks in a set of machine learning classification tasks. This investigation allowed us to compare between the predictive power of manual vowel onset annotations and acoustic landmarks, as well as ensure that the signal event detection parameters (e.g., minimum inter-peak interval), which were optimised using the development data set, have utility in a data set consisting of unseen speakers' data. To produce predictors for the classification tasks, inter-landmark and inter-vowel onset intervals (ms) were first calculated. For each speech window, this gave us a vector from which the following statistical parameters were derived: count, mean, mean absolute deviation, coefficient of variation, median; median absolute deviation, and coefficient of quartile variation. The binary speech rhythm classification tasks were

(1) English versus Mandarin, using data from all speaking conditions, with *article*/*poem* and *solo*/*joint* as additional predictors;

https://doi.org/10.1121/10.0009844

(2) reading versus spontaneous, using data from article reading and spontaneous speech trials, with *language* as an additional predictor;
(3) solo versus joint, using data from article and poem reading trials, with *language* and *article*/*poem* as additional predictors; and
(4) articles versus poems, using data from article and poem reading trials, with *language* and *solo*/*joint* as additional predictors.

The classifications were performed using a SVM [function fitcsvm( ) in MATLAB (The MathWorks, Natick, MA)], which is a type of supervised machine learning algorithm that determines the optimum placement of a decision boundary such that the margin or distance between observations belonging to each class is maximised (Boser *et al.*, 1992). Well-suited to multidimensional datasets like the current one, SVMs use a mapping function to transform data from input space into data in feature space in search of between-class linear separability. Here, each of the four tasks was performed ten times using fivefold cross-validation on 80% of the data with 20% held for testing predictions. Input features were standardised and SVM hyperparameters were automatically optimised in an iterative process. We report average test accuracies across the ten runs.

III. RESULTS

A. Vowel onset estimation

First, to gain a sense of how similar the linear acoustic features were to one another, we calculated their mean correlation over the windowed speech. Given that the speech envelopes were derived from the same speech data, we should expect to see reasonably high mean r values if the differences between algorithms are trivial. The closest linear similarity was between envelope 1, a signal rectification envelope, and envelope 2, the Hilbert transform envelope (mean $r = 0.94$). By comparison, the correspondences between these envelopes and envelope 3, which filters the broadband signal into "loudness contours" based on a perceptual scale (Oganian and Chang, 2019; Zwicker *et al.*, 1979), are much lower (envelope 1 mean $r = 0.58$, envelope 2 mean $r = 0.54$). Envelope 4, which bandpass filters the broadband speech signal between [400,4000] Hz, has slightly stronger statistical relationships with envelope 1 (mean $r = 0.65$) and envelope 2 (mean $r = 0.60$), but its closest correlation is with envelope 4 (mean $r = 0.81$). The GTCC 1 produced more moderate correlations with the speech envelopes (mean r range = [0.55 0.67]). A heat map depicting these results is plotted in Fig. 3.

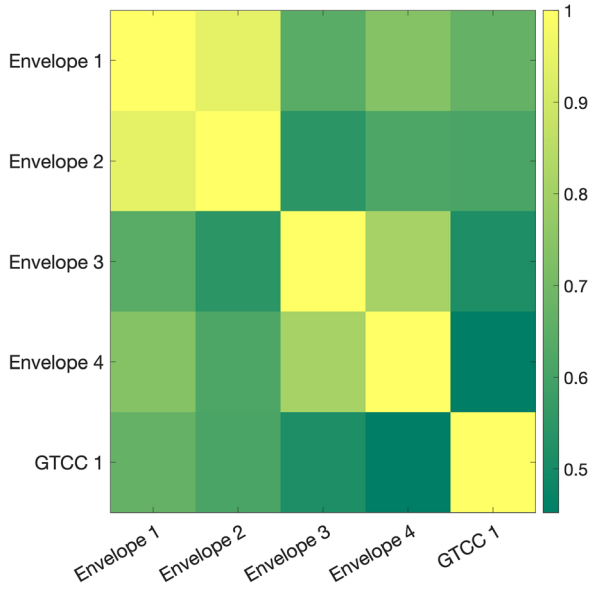

FIG. 3. (Color online) The mean Pearson's r correlation values between the acoustic features used for vowel onset estimation, averaged across speech windows.

1. Portion of matching windowed counts between manual annotations and acoustic landmarks

Turning to discrete events within the acoustic features, acoustic landmarks, our first objective was to maximise the portion of speech window data, where counts of landmarks fell within ±10% of corresponding counts of manual vowel onset annotations, by iterating through different algorithmic parameters. This was performed for each combination of feature and signal event. We retained the optimised parameters for each acoustic landmark. The results of this step are plotted in Fig. 4. A table of abbreviated results, showing the highest-ranked signal event for each feature, is given in Appendix A, Table V. Portions of matching counts by window varied substantially across acoustic landmarks and by speaking context. The lowest average rate for English was 0.21 (envelope 4/lower crossings) and 0.3 (envelope 4/mid-crossings) for Mandarin, and the highest average rates were tied at 0.56 (envelope 1/peaks; envelope 2/peaks; envelope 3/peaks) for English and 0.67 (envelope 3/peaks) for Mandarin. This first step allowed us to prune the search space as it were and focus only on acoustic landmarks with counts that more closely matched those of manual annotations.

2. Vowel estimation score

For each optimised acoustic landmark from the previous step, we calculated a weighted score that aggregated the portions of unmatched annotations and unmatched landmarks (i.e., false negatives and false positives, respectively) with the median error (ms) between paired landmarks and annotations. We calculated this vowel estimation score within speaking condition and then ranked the acoustic landmarks by their mean score across condition, within language, choosing the best three for English and best three for Mandarin to take forward. These results are shown in Appendix A, Table VI, and their vowel estimation scores and the algorithmic parameters used to find the signal events are shown in Appendix A, Table VII. All but one of the six selected acoustic landmarks was produced using envelope 3.

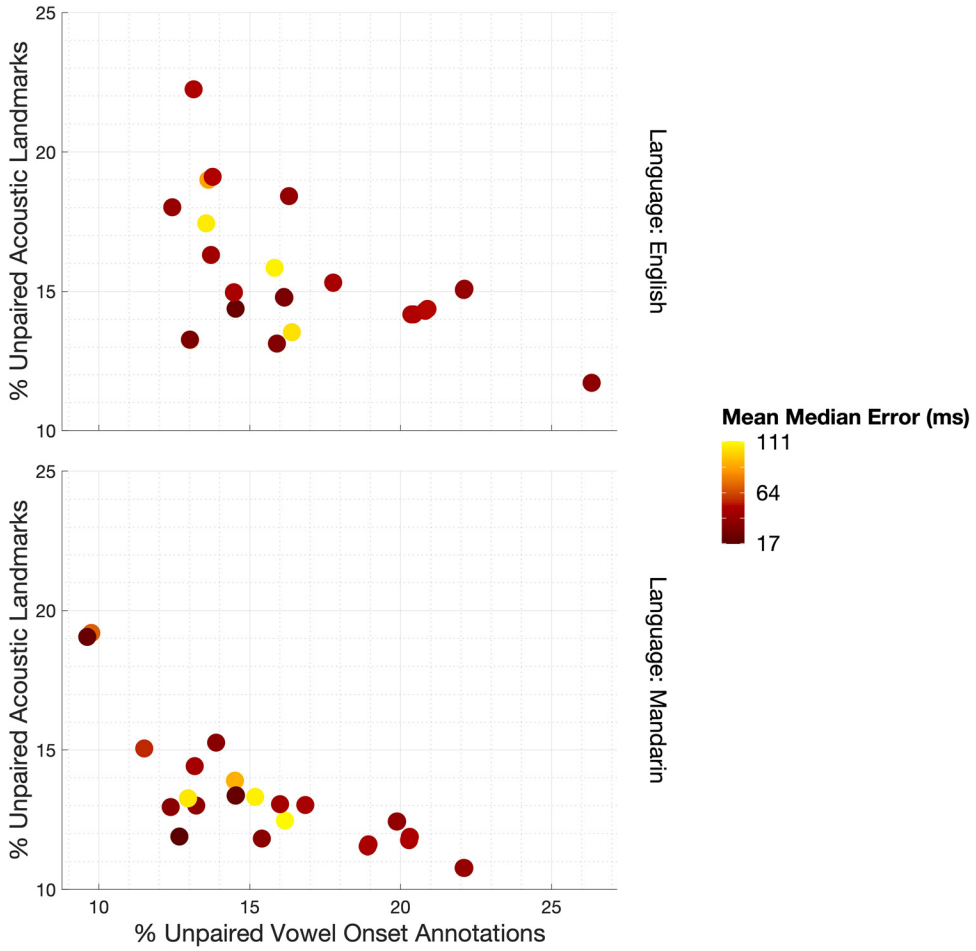

FIG. 4. (Color online) The acoustic landmarks, optimised with the aim of windowed counts falling within ±10% of the corresponding counts of manual vowel onset annotations. On the *Y* axis, unpaired acoustic landmarks refer to the portion of acoustic landmarks for which there was no mutual two-way closest match with manual vowel onset annotations. On the *X* axis, unpaired vowel onset annotations refer to the portion of manual annotations for which there was no mutual two-way closest match with acoustic landmarks. Each point represents the mean value taken over the three speaking conditions. The marker shade indicates the mean window median error (ms) between paired acoustic landmarks and vowel onset annotations.

For English and Mandarin, the overall highest ranked acoustic landmark, according to the vowel estimation score, is envelope 3/peaks in the first derivative, which is the same landmark that is described by Oganian and Chang (2019), confirming the findings reported from their experimental corpus. Table I shows the current results, broken down into speaking condition, for this landmark. The results are separate for the 80% of development data used to select the landmarks and the 20% of development data that were held out to confirm. Examples of the overall best performing landmark, plotted with the relevant speech envelope and raw wave form, are given for English and Mandarin in Figs. 5 and 6, respectively.

The plots displaying the individual components of the vowel estimation score for the top-ranked acoustic landmarks are shown in Appendix A, Figs. 11 and 12.

Irrespective of language, spontaneous speech is associated with larger median errors between manual annotations and acoustic landmarks (mean = 9.6 ms) in comparison to articles and poem reading (mean = 7.5 ms). Spontaneous speech also generates a much higher rate of unpaired annotations (mean = 14%) than articles and poem reading (mean = 6%). Given the marked differences between read and spontaneous speech, this is not surprising as spontaneous speech can be characterised by the shortening or outright dropping of vowels prescribed in the written form (Howell and Kadi-Hanifi, 1991). Taken together, it appears that diverse linguistic contexts elicit differing results from automatically generated speech features and/or events, even within the same speaker. As such, it may not be that the "temporal envelope of speech […] corresponds to the syllabic rhythm of speech" (Ding and Simon, 2014) consistently, at least not as far as syllables are defined by phonetic theory (Cummins, 2012b).

3. Correlations between the statistics of the time series derived from manual annotations and acoustic landmarks

We also compared various inter-event interval statistical parameters between the selected landmarks and vowel onset annotations on a window by window basis. The goal of this analysis was to compare between distributions of

https://doi.org/10.1121/10.0009844

TABLE I. The single best annotation similarity scoring acoustic landmark for the estimation of manual vowel onset annotations in both English and Mandarin was envelope 3/peaks in the first derivative. Presented here are the associated mean portions unpaired annotations and landmarks (i.e., annotations for which there were no mutual two-way closest matching landmarks and vice versa) and mean median error (i.e., Euclidean distance between paired annotations and landmarks) in milliseconds. The results are displayed separately for the 80% of the development dataset that was used to determine the vowel estimation score and the 20% of the dataset that was used to confirm the findings.

English—Envelope 3/peaks in the first derivative								
	Development				Held out			
			Mean portion unpaired				Mean unpaired	
Speaking condition	Count paired	Median error (ms)	Landmarks	Annotations	Count paired	Median error (ms)	Landmarks	Annotations
Articles	1819	8.16	0.06	0.1	447	7.32	0.09	0.08
Poems	1837	7.58	0.1	0.05	461	7.41	0.10	0.07
Spontaneous	2253	9.76	0.1	0.12	556	9.68	0.08	0.12
Mandarin—Envelope 3/peaks in the first derivative								
	Development				Held out			
			Mean portion unpaired				Mean unpaired	
Speaking condition	Count paired	Median error (ms)	Landmarks	Annotations	Count paired	Median error (ms)	Landmarks	Annotations
Articles	2510	7.74	0.05	0.04	630	8.73	0.04	0.04
Poems	1573	6.62	0.04	0.06	404	6.63	0.03	0.05
Spontaneous	2580	9.52	0.05	0.16	626	9.93	0.06	0.17

acoustic landmarks and manual annotations calculated from the same windowed speech, thereby permitting an objective assessment of the similarity in time series between techniques. As the same acoustic landmark was chosen first across both languages, this meant there were a total of five unique landmarks from which to calculate inter-event interval statistics to be correlated with those of manual annotations. The complete results appear in Appendix A, Table VIII. In general, Mandarin is associated with higher and less variable mean r values than English. The overall highest average correlation is between Mandarin annotations and envelope 3/peaks in the first derivative (mean $r = 0.59$), which is also the acoustic landmark with the highest vowel estimation score across both languages. On the other hand, this same landmark is relatively poorly correlated with English annotations (mean $r = 0.45$). The highest values for English were from envelope 3/peaks (mean $r = 0.51$). In any case, these moderate mean r values suggest that although the

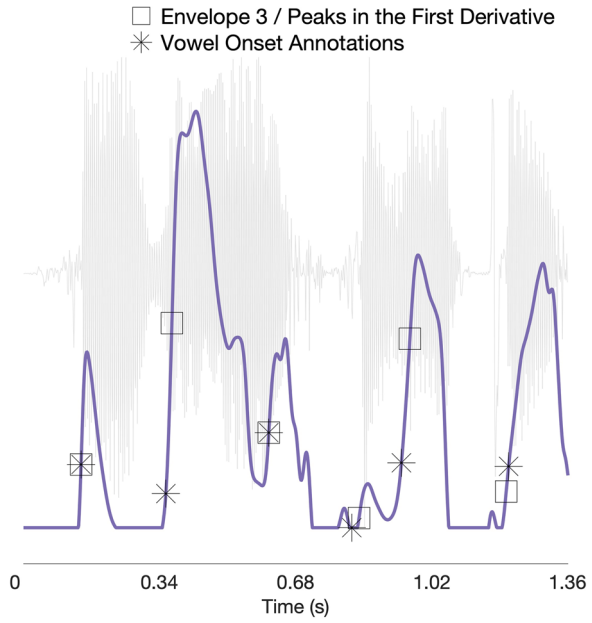

FIG. 5. (Color online) The highest-ranked acoustic landmark in English based on the vowel estimation score, plotted against the acoustic feature and raw speech wave form.

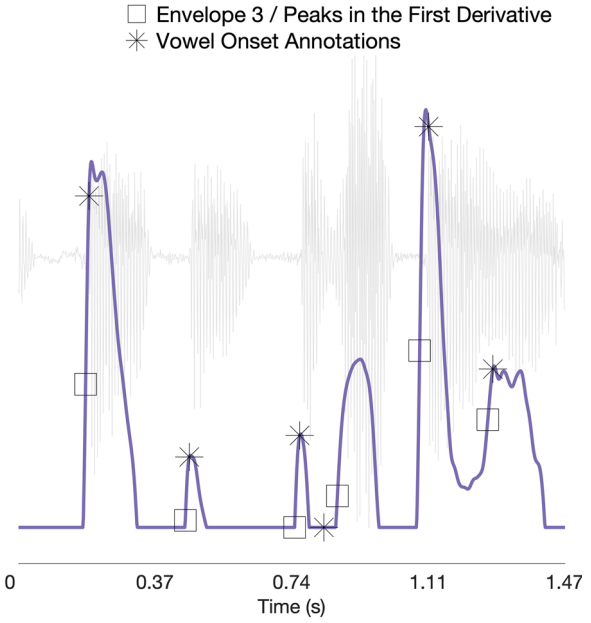

FIG. 6. (Color online) The highest-ranked acoustic landmark in Mandarin based on the vowel estimation score, plotted against the acoustic feature and raw speech wave form.

distributions of inter-landmark and inter-annotation intervals do correspond within speech windows, there is also considerable divergence.

Because nearly all of the landmarks selected by this stage were produced using the same acoustic feature, which was envelope 3, we reran the Pearson's correlations using the same signal event (peaks in the first derivative) but with each of the different acoustic features (GTCC 1, envelopes 1–4). This allowed us to gain some sense of the consequences of choice of envelope when all other parameters are held constant. The results are shown in Fig. 7 and Appendix A, Table IX. It is apparent that the different envelope techniques produce heterogeneous distributions of inter-landmark intervals, although the windowed counts of acoustic landmarks generally correspond well to those of manual vowel onset annotations across envelopes and languages. When averaged across the different statistics, the overall lowest mean *r* value was for envelope 2, which was the Hilbert transform (0.26). The highest mean *r* for English was envelope 3 (0.45). With regard to Mandarin speech, the lowest mean *r* was also envelope 2 (0.43), and the highest mean *r* was also envelope 3 (0.59).

B. The coordination of landmarks between speakers during joint speech

A higher density of landmarks across time will inherently bias toward smaller asynchronies, on average, between speakers engaged in joint speech. We, therefore, first selected for a vowel-like time scale by using the same acoustic landmarks that best matched the count per window of manual vowel onset annotations in the vowel onset estimation analysis, in addition to the joint speech-specific acoustic landmarks, consisting of each signal event combined with the GTCC 2–3 acoustic features. The results from this step are shown in Fig. 8. We searched for acoustic landmarks that minimise both unpaired observations and asynchrony (i.e., Euclidean distance in ms) between speakers. To this end, again, we employed a weighted score, the joint speech score, which accounted for mean median asynchronies and mean percent unpaired landmarks averaged across both speakers. The summary results for the highest ranking acoustic landmarks according to the joint speech score are shown in Appendix B, Table X. In both languages, there were landmarks tied by the joint speech score, indicating a similar degree of coordination between speakers for these acoustic landmarks. For brevity, we highlight envelope 3/peaks in the first derivative for English and GTCC 3/peaks for Mandarin, although both of these landmarks were similarly closely coordinated across both languages. Examples of the highlighted landmarks plotted against speech are given for English and Mandarin in Figs. 9 and 10, respectively. The plots displaying the individual components of the joint speech score for the top-ranked acoustic landmarks are shown in Appendix B, Figs. 13 and 14.

A breakdown by speaking condition of the results associated with these two landmarks is given in Table II. Surprisingly, there did not appear to be differences related to speaking condition despite poems being ostensibly rhythmically constrained and, in theory, more predictable in comparison to articles. Mandarin was, however, more closely

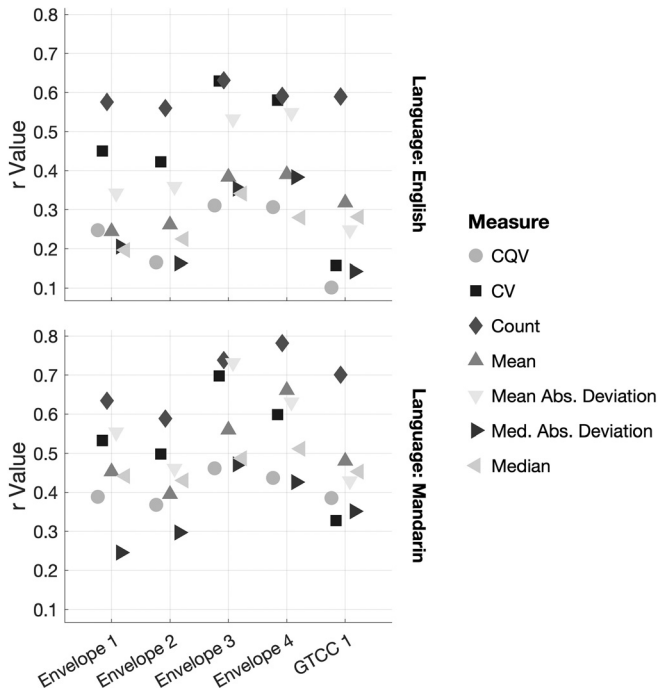

FIG. 7. Pearson's *r* correlation values between statistical parameters calculated from inter-acoustic landmark intervals and those calculated from corresponding manual vowel onset annotations in windowed speech. The same signal event, peaks in the first derivative, is used in combination with each of the different acoustic features used for manual vowel onset estimation. CQV, coefficient of quartile variation; CV, coefficient of variation.

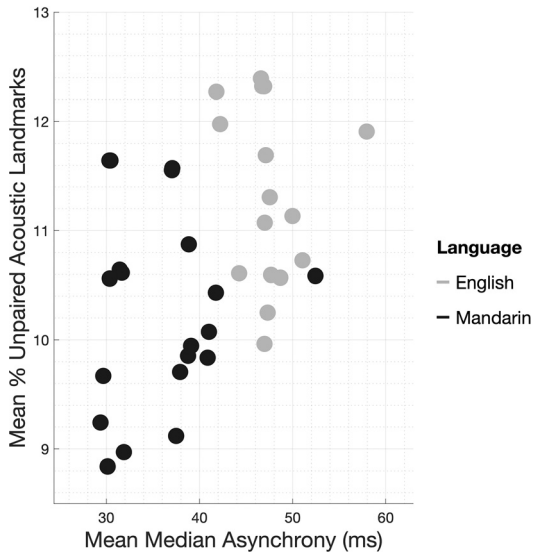

FIG. 8. The acoustic landmarks, optimised with the aim of windowed counts falling within ±10% of corresponding counts of manual vowel onset annotations. On the *Y* axis, unpaired acoustic landmarks refer to the portion of acoustic landmarks for which there was no mutual two-way closest match between speakers during joint speaking tasks. The *X* axis indicates the mean window median asynchrony (ms) between acoustic landmarks paired across speakers.

https://doi.org/10.1121/10.0009844

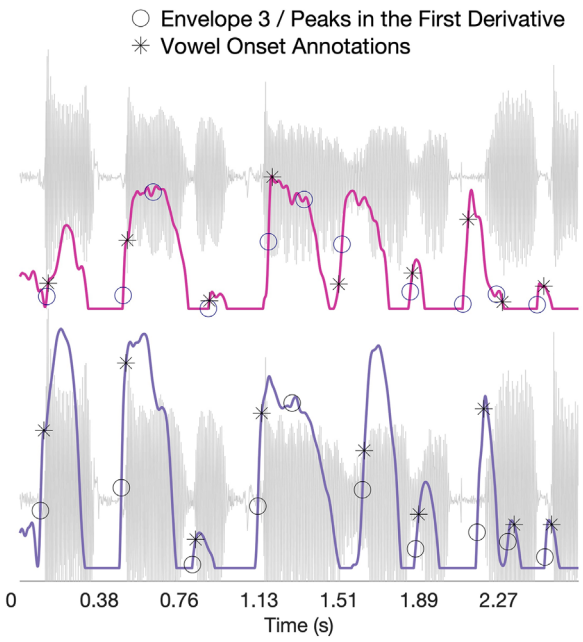

FIG. 9. (Color online) One of the tied highest-ranked acoustic landmarks in English based on the joint speech score, plotted against the raw speech wave form. The upper and lower panels depict each speaker's individual time series during the joint speaking task.

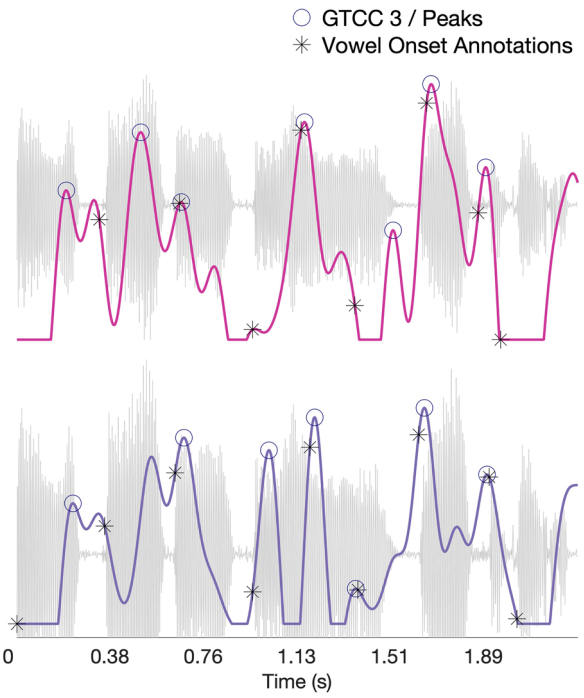

FIG. 10. (Color online) One of the tied highest-ranked acoustic landmarks in Mandarin based on the joint speech score, plotted against the raw speech wave form. The upper and lower panels depict each speaker's individual time series during the joint speaking task.

coordinated than English both in terms of asynchrony as well as the portion of landmarks that could not be paired between speakers. Based on visual inspection of GTCC 3/peaks, its position is also generally close to vowel onsets, although not as consistently as envelope 3/peaks in the first derivative.

For reference, we also ran the same joint speech analysis using manual vowel onset annotations, and the results are shown in Table III. As with the acoustic landmark results, there did not appear to be any substantial difference between speaking conditions, but Mandarin vowel onsets were more closely coordinated than those in English. The inter-speaker asynchrony values for manual annotations compared to acoustic landmarks were very similar overall, for example, a median of about 35 ms for acoustic landmarks versus 32 ms for annotations during English article

TABLE II. For both English and Mandarin, multiple acoustic landmarks tied for highest joint speech score. Presented here are a selection of results for envelope 3/peaks in the first derivative for English and GTCC 3/peaks for Mandarin only. Mean median asynchrony refers to the Euclidean distance separating landmarks paired between speakers in milliseconds. Mean portion unpaired refers to landmarks for which there were no mutual two-way closest matches between speakers. The results are displayed separately for the 80% of the development dataset that was used to determine the joint speech score and the 20% of the dataset that was used to confirm the findings.

English—Envelope 3/peaks in the first derivative										
	Development					Held out				
		Asynchrony (ms)		Portion unpaired			Asynchrony (ms)		Portion unpaired	
Speaking condition	Count paired	Median	Interquartile range	Mean	SD	Count paired	Median	Interquartile range	Mean	SD
Articles	1652	34.81	49.31	0.08	0.04	373	31.86	40.96	0.08	0.04
Poems	974	32.3	48.38	0.08	0.04	234	36.01	42.45	0.09	0.05
Mandarin—GTCC 3/peaks										
	Development					Held out				
		Asynchrony (ms)		Portion unpaired			Asynchrony (ms)		Portion unpaired	
Speaking condition	Count paired	Median	Interquartile range	Mean	SD	Count paired	Median	Interquartile range	Mean	SD
Articles	1970	22.73	31.19	0.06	0.04	485	21.22	32.24	0.07	0.03
Poems	861	22.77	30.18	0.06	0.04	234	21.92	31.48	0.05	0.03

TABLE III. The coordination of manual vowel onset annotations between speakers for English and Mandarin. Mean median asynchrony refers to the Euclidean distance separating annotations paired between speakers in milliseconds. Mean portion unpaired refers to annotations for which there were no mutual two-way closest matches between speakers.

English					
		Asynchrony (ms)		Portion unpaired	
Speaking condition	Counts paired	Median	Interquartile range	Mean	SD
Articles	626	32	44	0.05	0.03
Poems	643	32	41	0.03	0.03
Mandarin					
		Asynchrony (ms)		Portion unpaired	
Speaking condition	Counts paired	Median	Interquartile range	Mean	SD
Articles	863	25	34	0.02	0.02
Poems	557	23	28.25	0.02	0.02

reading. The equivalent values for Mandarin were 23 ms for acoustic landmarks versus 25 ms for annotations.

C. Application of acoustic landmarks in SVM classification of speech rhythm

In Secs. III A and III B, we analysed acoustic landmarks, first, from the perspective of trying to find the best match for manual vowel onset annotations, and, second, in terms of determining whether some acoustic landmarks are more closely coordinated between speakers than others. From each of these two steps, we selected three acoustic landmarks per language to apply in a proof of concept series of classification tasks using SVM. To this end, we calculated distributional parameters derived from windowed inter-landmark intervals as predictors, including count, median, and coefficient of variation, among others. This analysis was performed with the data from speakers who were held out entirely up until this point (see Appendix C, Table XI for a summary of observation counts by data set and task).

We report in Table IV the results for the SVM speech rhythm classification tasks by manual vowel onset annotation and selected acoustic landmarks. In general, mean classification accuracy was well above chance in most of the tasks for all of the selected acoustic landmarks, although inter-task performance could be quite variable. As in the vowel onset estimation and joint speech analyses, envelope 3/peaks in the first derivative (mean accuracy = 66%) and GTCC 3/peaks (mean accuracy = 65%) had the best average performance among the acoustic landmarks. For comparison, envelope 1/peaks in the first derivative produced a mean accuracy of 59%, a drop in 7% by changing the form of speech envelope. Manual vowel onset annotations were associated with superior classification performance (mean accuracy = 71%), indicating a level of error, loss, or difference in information for automatic techniques in comparison to manually estimated time series. Nonetheless, these classification rates, produced using uncorrected automatic techniques, demonstrate that variation in the speech timing series can be captured to an extent comparable with phonetic annotation.

IV. DISCUSSION

In the speech sciences, analysis has traditionally relied on expert knowledge and careful annotation by hand to study linguistically defined phenomena, such as syllables. More recently, some researchers, particularly those working in psychology and neuroscience, have used a variety of approaches to extract relevant features from the raw acoustic speech signal to estimate the speech envelope. These different signals can, in turn, be combined with various discrete events identified within those features. Although neuroscientists typically invoke concepts taken from linguistics, for example, when making claims about the brain's ability to track the syllabic time series, there is limited empirical evidence to support the interchangeability of phonetic ground truth with engineered acoustic features. Moreover, despite the variability in methods reported across the literature, we found that the choice of pipeline can produce diverse acoustic landmarks that vary in their similarity to manual vowel onset annotations. Choosing the vowel onset, a phonetic concept that is closely associated with *p*-centres,

TABLE IV. The results for the SVM speech rhythm classification tasks by manual vowel onset annotation and acoustic landmark.

Acoustic landmark		Accuracy				
Feature	Signal event	English versus Mandarin	Reading versus spontaneous	Solo versus joint	Articles versus poems	Mean
Manual vowel onset annotations		0.70	0.72	0.72	0.68	0.71
Envelope 3	Peaks in the first derivative	0.61	0.69	0.65	0.69	0.66
Envelope 3	Mid-crossings	0.64	0.69	0.68	0.59	0.65
GTCC 3	Peaks	0.64	0.69	0.65	0.63	0.65
Envelope 3	Lower crossings	0.64	0.67	0.68	0.56	0.64
Envelope 3	Peaks	0.56	0.71	0.65	0.62	0.63
GTCC 3	Peaks in the first derivative	0.63	0.68	0.64	0.54	0.62
GTCC 1	Peaks in the first derivative	0.59	0.61	0.65	0.64	0.62
Envelope 1	Peaks in the first derivative	0.57	0.62	0.63	0.54	0.59

as our ground truth, we systematically compared different automatically generated acoustic landmarks with manual annotations. We also explored how closely and consistently speakers coordinated different acoustic landmarks when speaking together during joint speech tasks. Finally, we used a subset of acoustic landmarks in a proof of concept, which was a set of speech type classification tasks. Using statistics extracted from temporal information alone, we found reasonably good classification of diverse speaking styles, but the accuracy of landmarks was reduced in comparison with manual annotation. Our results indicate that acoustic landmarks do not perfectly replicate the syllabic or inter-vowel time series, and specific algorithmic and parametric choices produce quantifiably different results. Hence, researchers who draw on phonetic theory in motivating behavioural or neural experiments should bear this in mind when choosing whether or not to manually annotate, and if not, when determining the appropriate automatic technique. We cover these results in more detail in the following discussion.

A. Vowel onset estimation

Our results show that any one algorithm can generate varying results not only with regard to the particular language under study but within languages as well. For instance, vowel estimation was more successful in Mandarin than in English overall, yet even the best overall acoustic landmark left up to mean 16% of manual vowel annotations unpaired in Mandarin spontaneous speech. In contrast, this same landmark produced just 4% unpaired annotations in Mandarin article reading. Our finding here speaks to the work by Schuppler (2017), who found that machine phonemic classification is also hampered in spontaneous speech due to larger acoustic overlap between classes in comparison to carefully read speech. Instead of thinking of spontaneous speech as the noisier version of read speech, Schuppler (2017) argues that the unique properties of spontaneous speech be considered early in the development of methodologies. Similarly, rather than ponder whether an alternative algorithm might have done a better job, we propose that higher rates of "missed" or unpaired vowel onset annotations may indicate that the map of the written system is a poor fit to the territory of spontaneous speech. In other words, phonetically determined units, such as syllables, may be simply inappropriate for the analysis of rhythm in spontaneous speech. This complexity is compounded by the inherent subjectivity of manual annotation; although we took pain to ensure consistency across the current corpus, the expert knowledge and judgment calls required in phonetic analysis present considerable challenges when interpreting results across studies and research groups. Moreover, there remains controversy within linguistics concerning the exact boundaries between consonants and vowels (Francis *et al.*, 2003) let alone syllables themselves (Goslin and Frauenfelder, 2001; Zec, 2007). For this reason, choosing a consistent acoustic landmark rather than linguistic token may actually be of benefit to speech rhythm research more broadly, especially in natural, coarticulated speech, where phonetic truisms may be more liable to break down. That stated, should a researcher nonetheless wish to emulate vowel onsets specifically, it is possible that even a perfunctory visual review of acoustic landmarks could be enough to bring the current results closer in line to those of manual annotations given that we did not apply any corrective procedure to the landmarks here.

Among the different engineered features that we tested, the envelope-extracting method described in Oganian and Chang (2019) and Schotola (1984), envelope 3, produced results most similar to manual vowel onset annotations. This psychoacoustically informed feature was robust in approximating vowel onsets across two unrelated languages and diverse speaking contexts. In comparison, envelope 2, which uses the Hilbert transform of the broadband speech signal, performed more inconsistently. This technique has drawn criticism for producing distorted or inaccurate modulation frequencies (Schimmel and Atlas, 2005). It could be that incorporating improvements proposed by Schimmel and Atlas (2005) and others in the signal processing community would have produced better results, but our intention here was to emulate practices typical of the speech rhythm and neuroscience literatures. In terms of the five discrete events within the signals that we examined, peaks in the first derivative (or rate of change) of the signal tended to best approximate the location of manual vowel onset annotations. This signal event, when combined with envelope 3, produces the same acoustic landmark that was found by Oganian and Chang (2019) to be specifically encoded by a defined region of the STG during speech perception. Taken together, these findings show that the choice of feature and discrete events are nonarbitrary for approximating the location vowel onsets within the signal despite the variety of reported procedures.

B. Joint speech

In comparison to manual annotations, landmarks were associated with slightly higher rates of unpaired events between speakers during joint speech, although median asynchrony values were roughly on par. We found that the acoustic landmark that performed best in vowel onset estimation, envelope 3/peaks in the first derivative, was also tightly synchronised between speakers of both English and Mandarin, aligning with the behavioural research that approximates *p*-centres to this same phonetic anchor (Barbosa *et al.*, 2005; Rathcke *et al.*, 2021; Scott, 1998). GTCC 3/peaks, however, was roughly equally well coordinated in joint speech. It should be noted that we found GTCC 3/peaks to mainly occur closely around envelope 3/peaks in the first derivative in our corpus. GTCC 3/peaks may, therefore, similarly channel vowel onsets, although we did not test this particular acoustic landmark in the vowel estimation analysis. Future work could examine these different, GTCC-based acoustic landmarks in the context of more in-depth phonetic analysis and behavioural rhythm perception and production experiments.

TABLE V. The portion of matching counts, by window, between manual vowel onset annotations and acoustic landmarks. The single highest-ranked signal event is presented for each feature by language. Portion matching counts refer to the portion of windows where the count of landmarks falls within ±10% of the corresponding count of manual annotations for that window.

	Acoustic landmark		Portion matching counts			
Language	Feature	Signal event	Articles	Poems	Spontaneous	Mean
English	GTCC 1	Bases of peaks	0.53	0.58	0.44	0.52
English	Envelope 1	Peaks	0.57	0.59	0.52	0.56
English	Envelope 2	Bases of peaks	0.61	0.61	0.46	0.56
English	Envelope 3	Peaks	0.55	0.67	0.45	0.56
English	Envelope 4	Peaks	0.47	0.61	0.46	0.51
Mandarin	GTCC 1	Bases of peaks	0.65	0.59	0.48	0.57
Mandarin	Envelope 1	Peaks	0.74	0.72	0.43	0.63
Mandarin	Envelope 2	Bases of peaks	0.71	0.72	0.38	0.6
Mandarin	Envelope 3	Peaks	0.76	0.82	0.44	0.67
Mandarin	Envelope 4	Peaks	0.77	0.83	0.39	0.67

C. Speech classification

Although the landmarks did not match the speech rhythm classification performance of manual annotations (mean = 71%), we were still able to achieve mean accuracy as high as 66% in the held-out corpus with completely novel speakers using temporal information alone. In light of the time and resources required for manual annotations, our results demonstrate that some—if not all—engineered features have the capability to produce meaningful insight into speech timing, they are not a perfect replication of phonetic analysis. With regard to the individual classification tasks, we noticed that the acoustic landmark closest to vowel onsets, envelope 3/peaks in the first derivative, was less able to discriminate between English and Mandarin speech (accuracy = 61%) in comparison to annotations (accuracy = 70%). We suspect this is related to vowel shortening and coarticulation effects in English, which are more easily identified in the manual annotations rather than in acoustic landmark data.

This discrepancy could indicate the impracticality or trade-off in using automatic extraction techniques. On the other hand, cross-linguistic similarity between acoustic landmarks may tell us more about language-invariant aspects of speech perception (Räsänen *et al.*, 2018). For example, the variability of inter-vowel intervals in languages like English is often contrasted with the simpler syllabic structures of Mandarin (Lin and Wang, 2007). But this seeming dichotomy may be amplified by expert knowledge of written forms and linguistic theory. Indeed, Oganian and Chang (2019) found similar neural responses to English and Mandarin stimuli despite the differing linguistic backgrounds of their participants. Future work should confirm whether the acoustic landmarks examined in the current study evoke comparable brain activity across speakers of different languages.

D. Limitations

As we have discussed, an inherent limitation to the current work is that manual annotation is ultimately subjective, and we cannot exclude that the specific manner in which we identified vowels and determined the precise moment of onset has some impact on the generalisability of the current results. That said, the acoustic landmark that we observed to be most similar to vowel onsets (envelope 3/peaks in the first derivative) is the same as that reported in Oganian and Chang (2019), who confirmed their results in Mandarin and Spanish speech data sets, in addition to the TIMIT corpus in English (Garofolo *et al.*, 1993). Although we similarly

TABLE VI. The best-scoring three acoustic landmarks for English and Mandarin according to the vowel estimation score and their associated mean portions unpaired annotations and landmarks (i.e., annotations for which there were no mutual two-way closest matching landmarks and vice versa) and mean median error (i.e., Euclidean distance between paired annotations and landmarks) in milliseconds.

					Portion unpaired landmarks			Portion unpaired annotations		
			Error (ms)			95% confidence intervals			95% confidence intervals	
Language	Feature	Signal event	Median	Interquartile range	Mean	Lower	Upper	Mean	Lower	Upper
English	Envelope 3	Peaks in the first derivative	8.48	12.61	0.1	0.09	0.11	0.1	0.09	0.11
English	Envelope 1	Peaks in the first derivative	23.35	26.23	0.1	0.1	0.11	0.13	0.12	0.14
English	Envelope 3	Peaks	39.3	36.7	0.08	0.07	0.09	0.11	0.1	0.12
Mandarin	Envelope 3	Peaks in the first derivative	8.29	10.61	0.07	0.06	0.08	0.09	0.08	0.1
Mandarin	Envelope 3	Lower crossings	7.59	10.89	0.07	0.07	0.08	0.12	0.11	0.13
Mandarin	Envelope 3	Mid-crossings	7.55	10.92	0.07	0.06	0.08	0.12	0.11	0.13

https://doi.org/10.1121/10.0009844

TABLE VII. The three highest ranked acoustic landmarks for English and Mandarin according to the vowel estimation score. This score combines the portions of unpaired landmarks and annotations with the median error (ms). Prominence, inter-peak interval, and tolerance are algorithmic parameters used to identify the signal events.

Language	Feature	Signal event	Prominence	Inter-peak interval (ms)	Tolerance	Vowel estimation score
English	Envelope 3	Peaks in the first derivative	0.1	115	—	0.71
English	Envelope 1	Peaks in the first derivative	0.1	120	—	0.61
English	Envelope 3	Peaks	0.1	105	—	0.53
Mandarin	Envelope 3	Peaks in the first derivative	0.1	115	—	0.78
Mandarin	Envelope 3	Lower crossings	—	—	4	0.72
Mandarin	Envelope 3	Mid-crossings	—	—	4	0.72

found this same landmark to be robust across language and speaking style, it is important to replicate the results within a wider variety of languages. On a related note, the algorithmic parameters we report for signal event detection (e.g., peak detection) were again optimised for vowel onsets specific to the current data. We attempted to control over-fitting by splitting our corpuses into speaker-independent development and testing data sets, including partitioning the former during acoustic landmark selection to further confirm that our choices were generalisable. However, we communicate these "optimum" parameters with the caveat that they are likely only trustworthy to a certain level of granularity, and further work across many diverse speech corpuses is needed before we can be satisfied of any universal best fit.

As we explained in the Introduction, we take no firm stance regarding the meaning or measure of speech rhythm. We have focused on syllable timing in the current work due to the considerable theoretical and experimental attention it has received, but there are many ways of expressing rhythm. In the context of musicking (Small, 1998), for instance, patterns in timing can emerge from dance steps, pitch groupings, and variation in loudness, accent, articulation, and timbre, in addition to other features. By comparison, in speech sciences and especially neuroscience, the intense focus on timing and intervals has possibly resulted in analogous qualities in speech rhythm being overlooked. Although the current paper does not break from this status quo, other aspects of prosody, such as pitch, should be incorporated into temporal frameworks (Teoh *et al.*, 2019; Vicenik and Sundara, 2013). Similarly, there is also no reason why speech rhythm researchers cannot combine the speech envelope and/or syllable time series with phonetic annotations, as well as other engineered features, such as the GTCC that we explored in the current work. Increasingly, approaches like this have been used to map the correspondence between acoustic stimuli and the neural response, for example, by combining the envelope and its derivative (Drennan and Lalor, 2019) or spectro-temporal acoustic features and phonetic information (Di Liberto *et al.*, 2018). Using band limited versions of the speech envelope, Bröhl and Kayser (2021) recently demonstrated that listeners' brains responded differently to relatively lower and higher frequency components, revealing spatial and functional nuances that could be distorted or even lost when only the broadband speech envelope is used. Hence, although we have primarily focused on vowel onsets and automatic methods to characterise them, we acknowledge that the richness of speech rhythm calls for holistic methods that better reflect its complexity.

TABLE VIII. The results for the Pearson's correlations between statistical parameters calculated using inter-landmark intervals and inter-annotation intervals on the same windowed speech data. The selection of acoustic landmarks was determined according to the vowel estimation score. For each parameter, r is first calculated by speaking condition and then aggregated by language. Mean r is aggregated across all parameters.

	Acoustic landmark		
Language	Feature	Signal event	Mean r
English	Envelope 3	Peaks	0.51
English	Envelope 3	Lower crossings	0.48
English	Envelope 3	Mid-crossings	0.48
English	Envelope 3	Peaks in the first derivative	0.45
English	Envelope 1	Peaks in the first derivative	0.32
Mandarin	Envelope 3	Peaks in the first derivative	0.59
Mandarin	Envelope 3	Lower crossings	0.59
Mandarin	Envelope 3	Mid-crossings	0.59
Mandarin	Envelope 3	Peaks	0.59
Mandarin	Envelope 1	Peaks in the first derivative	0.48

TABLE IX. The results for the Pearson's correlations between statistical parameters calculated using inter-landmark intervals and inter-annotation intervals on the same windowed speech data. The acoustic landmarks consist of the same signal event, peaks in the first derivative, coupled with each different feature used in the vowel onset estimation. r is first calculated by speaking condition and then aggregated by language. Mean r is aggregated across all parameters. Note that the algorithmic parameters used to detect signal events were held constant across the different envelopes for this analysis, and so the mean r may be slightly different from the equivalent acoustic landmark reported in Table VIII.

	Acoustic landmark		
Language	Feature	Signal event	Mean r
English	Envelope 1		0.32
English	Envelope 2		0.31
English	Envelope 3		0.45
English	Envelope 4		0.44
English	GTCC 1	Peaks in the first derivative	0.26
Mandarin	Envelope 1		0.46
Mandarin	Envelope 2		0.43
Mandarin	Envelope 3		0.59
Mandarin	Envelope 4		0.58
Mandarin	GTCC 1		0.45

E. Conclusion

The case for cortical speech entrainment rests on the quality and specificity of experimental stimuli and the materials derived therefrom. It has been found that synthetic vowels evoke a stronger envelope-following cortical response than their natural counterparts, potentially due to enhanced, stabler periodicity in the frequency domain (Van Canneyt *et al.*, 2020). Similarly, in the speech rhythm domain, we should also expect to see differences in neural responses between artificially periodic stimuli and naturalistic speech, especially speech that is spontaneously produced. Selecting the appropriate acoustic landmark to capture characteristics of speech rhythm across a variety of speaking contexts, therefore, constitutes an important step toward the greater task of improving ecological validity across the field at large (Alexandrou *et al.*, 2020). In light of the behavioural literature on *p*-centres (Rathcke *et al.*, 2021; Scott, 1998; Seifart *et al.*, 2018) and convergent data from neuroimaging (Oganian and Chang, 2019), we suggest that researchers interested in speech rhythm select envelope extraction techniques that best convey information about vowels, especially vowel onsets. Of the features that we examined in the current study, we found this to be envelope 3, undermining the assumption that different speech envelope techniques are "similar in general" (Ding *et al.*, 2017) or even "qualitatively identical" in practice (Oganian and Chang, 2019); however, we imagine that other approaches to envelope extraction that emulate cochlear filtering of the broadband speech signal may perform similarly, a speculation that should be empirically tested in future work. In the case that discrete events are of interest, rather than the continuous signal, we additionally recommend identifying peaks in the first derivative of envelope 3. Comparing the discrete and continuous time series may help answer the question of whether specific instances within the speech envelope and not the speech envelope uniformly drive the processing of speech rhythm (Ding and Simon, 2014). If this were indeed the case, it is particularly important for studies that investigate stimulus-brain correlation; for instance, a continuous signal could,

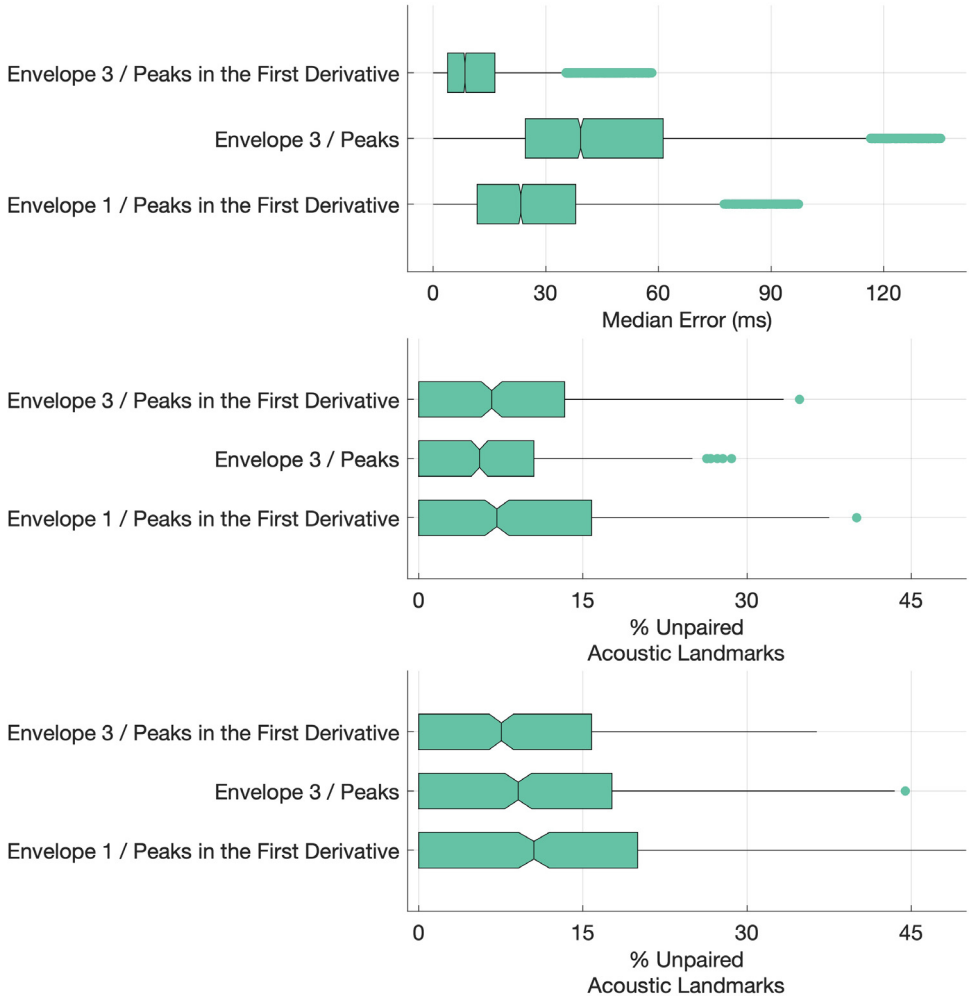

FIG. 11. (Color online) Box plots depicting the three highest-ranked acoustic landmarks in English based on the vowel estimation score. Unpaired acoustic landmarks refer to the portion of acoustic landmarks for which there was no mutual two-way closest match with manual vowel onset annotations. Unpaired vowel onset annotations refer to the portion of manual annotations for which there was no mutual two-way closest match with acoustic landmarks. The median error (ms) is the Euclidean distance between paired acoustic landmarks and vowel onset annotations. The notches indicate 95% confidence intervals for the median.

https://doi.org/10.1121/10.0009844

on the whole, appear to correspond well with neural activity when, in fact, crucial events within the signal are relatively poorly aligned with neural markers of interest or vice versa. Additionally, the current study illustrates the need to use varied and naturalistic stimuli as our results undermine the simple reading that the speech envelope is equivalent to phonetically defined syllables, which are themselves equivalent to the time series that should drive neural entrainment. At the least, it should not be assumed that different approaches to engineered features are more or less the same as phonetic annotation—nor to each other, for that matter, particularly across diverse speaking contexts. As a mechanism so central to human communication and cognition, more generally, rhythm has long deserved its due, and cooperation across and within disciplinary boundaries is required for progress to be made in elucidating its inner workings.

ACKNOWLEDGMENTS

A.D.M. is supported by the University College London Graduate and Overseas Research Scholarships. C. Q.C. is supported by the China Scholarship Council–University College London Joint Research Scholarship. The authors wish to thank Yulia Oganian and Sam Tilsen for sharing the code used to generate envelopes 3 and 4 in this study.

APPENDIX A: VOWEL ONSET ESTIMATION

See Tables V–IX for descriptive statistics concerning the vowel onset estimation analysis and Figs. 11 and 12 for box plots depicting the individual components of the vowel estimation score.

APPENDIX B: JOINT SPEECH

See Table X for descriptive statistics concerning the joint speech analysis and Figs. 13 and 14 for box plots depicting the individual components of the joint speech score.

APPENDIX C: SVM CLASSIFICATION OF SPEECH RHYTHM

See Table XI for class membership counts by classification task.

APPENDIX D: EXCERPTS

1. English article A

The Great Pyramid of Giza is the oldest monument of the seven wonders of the Ancient World. It is also the only one left standing. The structure is a marvel of human engineering, and its sheer size and scale rivals anything built within the last few hundred years. Its creation, however, has always been the subject of much debate among scholars because of its massive size and near perfect proportions.

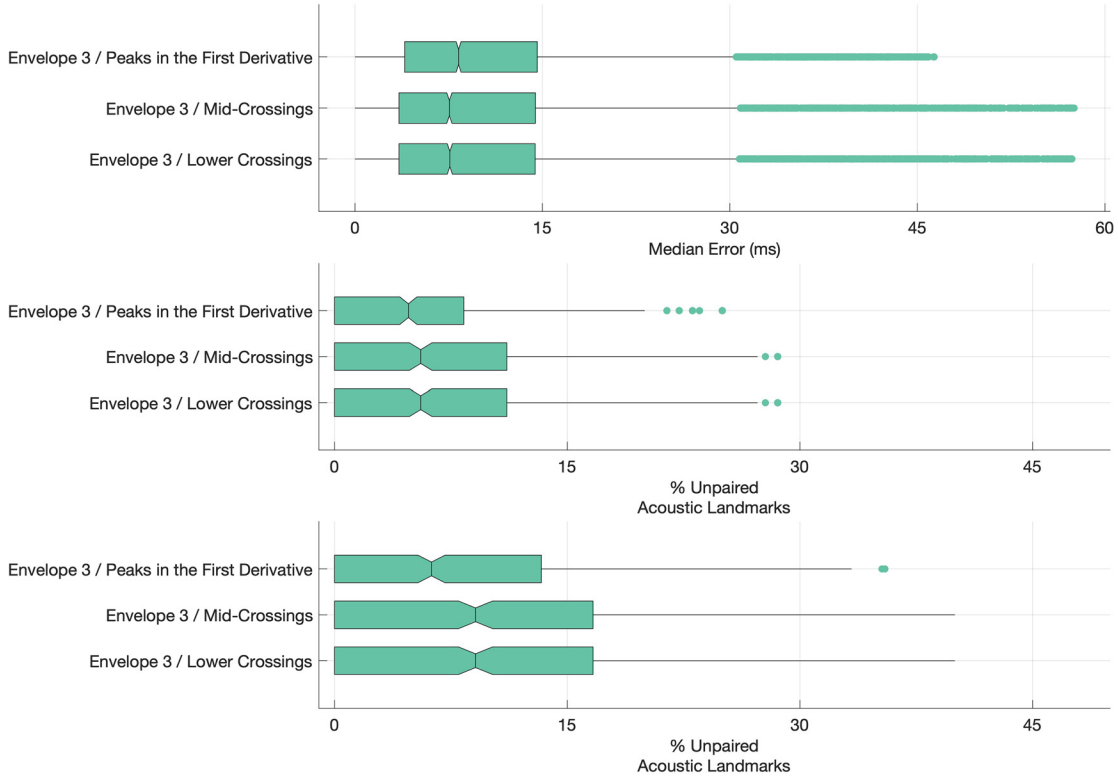

FIG. 12. (Color online) Box plots depicting the three highest-ranked acoustic landmarks in Mandarin based on the vowel estimation score. Unpaired acoustic landmarks refer to the portion of acoustic landmarks for which there was no mutual two-way closest match with manual vowel onset annotations. Unpaired vowel onset annotations refer to the portion of manual annotations for which there was no mutual two-way closest match with acoustic landmarks. The median error (ms) is the Euclidean distance between paired acoustic landmarks and vowel onset annotations. The notches indicate 95% confidence intervals for the median.

TABLE X. The best-scoring three acoustic landmarks for English and Mandarin according to the joint speech score and their associated mean portions unpaired landmarks (i.e., landmarks for which there were no mutual two-way closest matching landmarks between the two speakers) and mean median asynchrony (i.e., Euclidean distance between the speakers' paired landmarks) in milliseconds.

	Acoustic landmark				Asynchrony (ms)		Portion unpaired landmarks			
							95% confidence intervals			
Language	Feature	Signal event	Joint Speech score	Count paired	Median	Interquartile range	Mean	Lower	Upper	SD
Mandarin	GTCC 3	Peaks	0.80	2832	22.74	31.01	0.06	0.06	0.07	0.04
Mandarin	GTCC 3	Peaks in the first derivative	0.80	2816	24.45	36.01	0.06	0.06	0.06	0.03
Mandarin	Envelope 3	Peaks in the first derivative	0.80	2782	24.69	31.09	0.05	0.05	0.06	0.04
English	Envelope 3	Peaks in the first derivative	0.74	2622	34.17	48.46	0.08	0.07	0.08	0.04
English	GTCC 3	Peaks	0.74	2595	34.18	51.16	0.08	0.08	0.09	0.04
English	GTCC 1	Peaks in the first derivative	0.73	2595	36.08	53.14	0.08	0.07	0.09	0.04

Evidence suggests that 20 000 workers contributed to its construction and were even paid to do so. This would have required a great deal of organisation, accounting, and record keeping. The Egyptians were known for their excellent documentation.

2. English article B

Sir Frederick William Herschel discovered infrared light at the turn of the 19th century. Using a variety of coloured filters to view sunlight, he observed that some colours

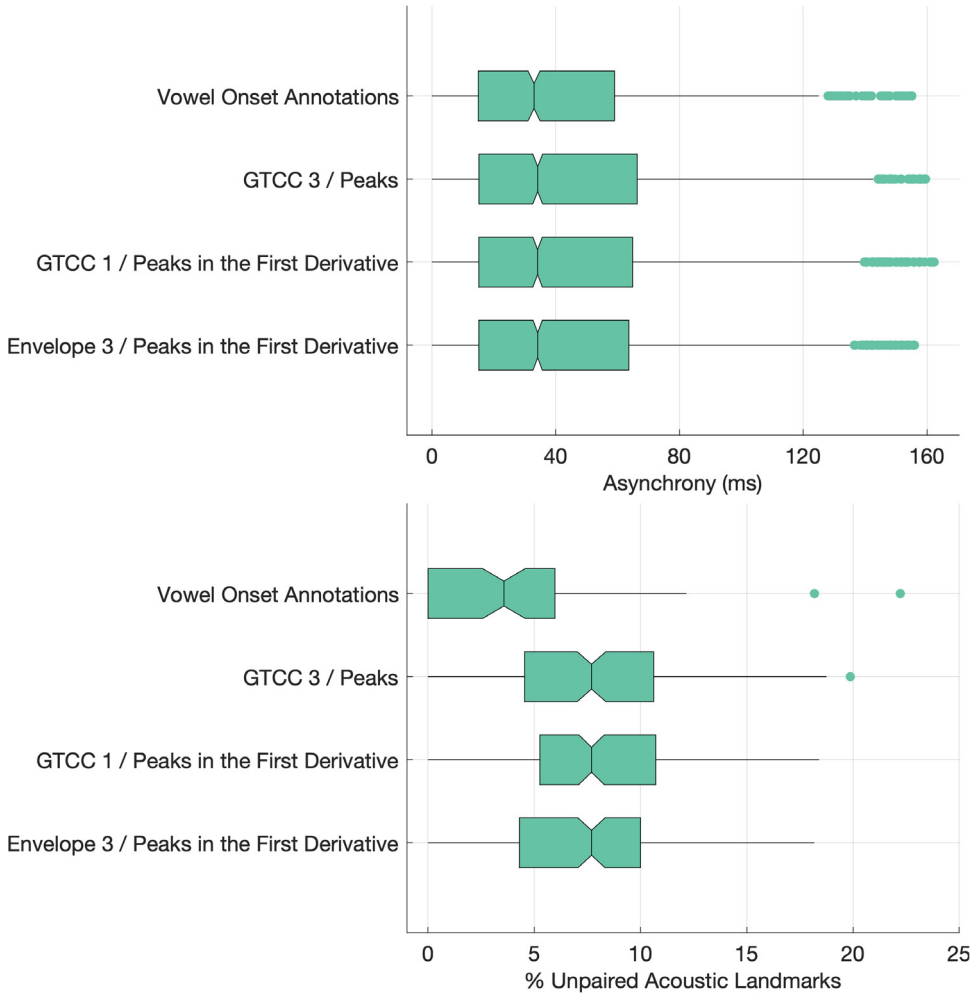

FIG. 13. (Color online) Box plots depicting the three highest-ranked acoustic landmarks in English based on the joint speech score. Unpaired acoustic landmarks refer to the portion of acoustic landmarks for which there was no mutual two-way closest match between speakers. The mean median error (ms) is the Euclidean distance between acoustic landmarks paired across speakers. For reference, the corresponding values for manual vowel onset annotations are also shown. The notches indicate 95% confidence intervals for the median.

https://doi.org/10.1121/10.0009844

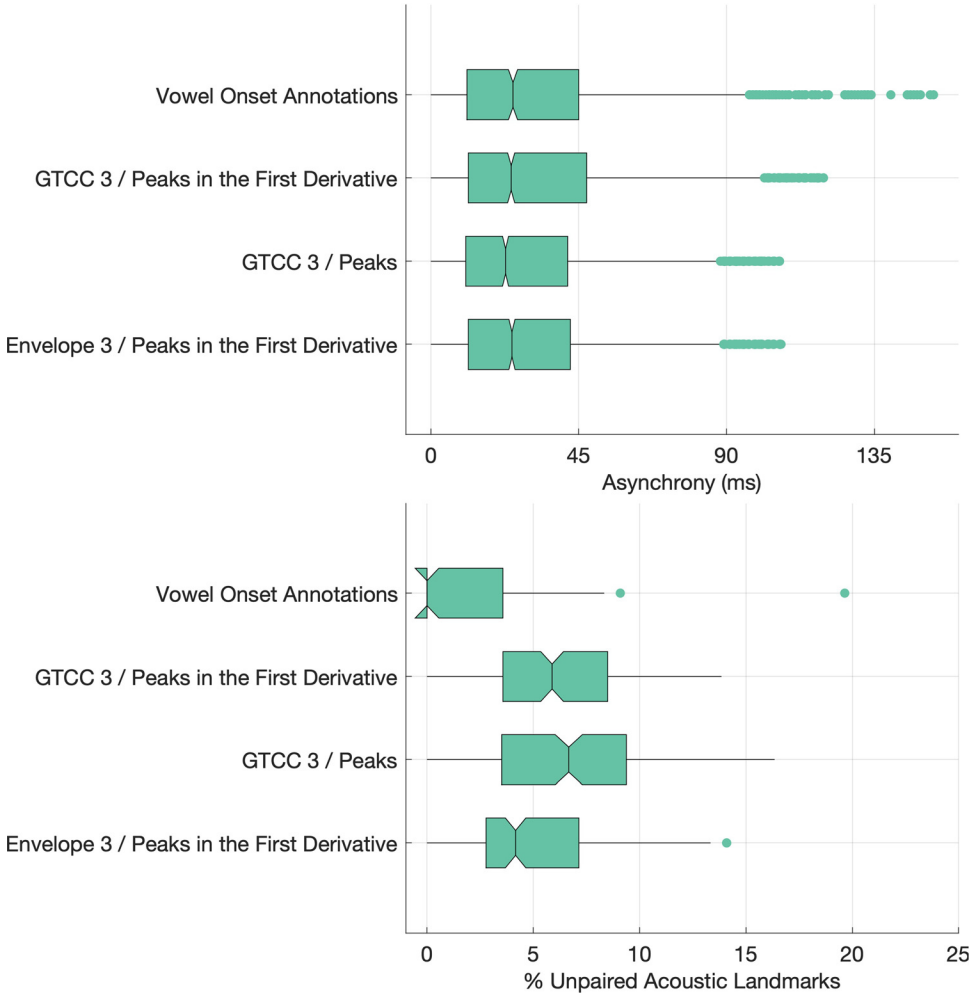

FIG. 14. (Color online) Box plots depicting the three highest-ranked acoustic landmarks in Mandarin based on the joint speech score. Unpaired acoustic landmarks refer to the portion of acoustic landmarks for which there was no mutual two-way closest match between speakers. The mean median error (ms) is the Euclidean distance between acoustic landmarks paired across speakers. For reference, the corresponding values for manual vowel onset annotations are also shown. The notches indicate 95% confidence intervals for the median.

allowed more heat to pass than others. Herschel hypothesised that the colours themselves may have produced different temperatures and set out to test his theory.

Directing sunlight through a glass prism to produce a spectrum, he measured the temperature of each colour with a thermometer. From the violet to red parts of the colour spectrum, the temperature increased. Herschel then decided to measure the temperature beyond the red-coloured light. This area had the highest temperature reading of all despite being invisible to the naked eye.

TABLE XI. Counts of windows from which statistical parameters were calculated to use as predictors in the speech rhythm classification tasks. The windowed data are from speakers whose recordings were entirely held out during the vowel onset estimation and joint speech analyses.

English versus Mandarin		Spontaneous versus reading	
	Count windows		Count windows
English	237	Spontaneous	130
Mandarin	258	Reading	79
Solo versus joint		Articles versus poems	
	Count windows		Count windows
Solo	156	Articles	196
Joint	209	Poems	169

3. English article C

The North Wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew, the more closely did the traveller fold his cloak around him, and at last the North Wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveller took off his cloak. And so, the North Wind was obliged to confess that the Sun was the stronger of the two.

4. English poem A. "Bed in Summer" by Robert Louis Stevenson

In winter I get up at night

And dress by yellow candle-light.
In summer, quite the other way,
I have to go to bed by day.
I have to go to bed and see
The birds still hopping on the tree,
Or hear the grown-up people's feet
Still going past me in the street.
And does it not seem hard to you,
When all the sky is clear and blue,
And I should like so much to play,
To have to go to bed by day?

5. English poem B. "The Thing With Feathers" by Emily Dickinson

"Hope" is the thing with feathers
That perches in the soul
And sings the tune without the words
And never stops – at all.
And sweetest in the gale is heard
And sore must be the storm
That could abash the little bird
That kept so many warm.
I've heard it in the coldest land
And on the strangest sea
Yet – never – in extremity,
It asked a crumb of me.

6. English prompts for spontaneous speech

What brings you to London?
What are some things you like about London?
What are some things you do not like about London?

Do you have a favourite restaurant or cuisine, and is it available in London?

Have you been to any of the parks in London? Do you have a favourite park?

If you met someone new to the city, what would you recommend they go to see or do?

Have you been to any tourist attractions (for example, the London Eye)? Did you enjoy it, or was it overrated?

Mandarin article A 吉萨金字塔是古代世界七大奇迹中最为古老的纪念碑。它也是唯一尚存的建筑物。它的结构是人类工程学的奇迹， 庞大的规模可与过去几百年来建造的任何建筑相媲美。然而，正因为它庞大的规模和接近完美的比例，如何创造它一直是学者们争论的主题。有证据表明两万名工人为建设它做出了贡献， 并为劳动获得了报酬。这需要大量的组织，会计和记录保存。埃及人以出色的文献记录而著称。

Mandarin article B 在19世纪初，弗雷德里克·威廉·赫歇尔爵士发现了红外光。 他使用各种彩色滤光片查看日光， 观察到某些颜色相比其他颜色允许更多的热量通过。赫歇尔假设颜色本身可能产生不同的温度， 并着手检验他的理论。他引导阳光通过玻璃棱镜产生光谱，然后用温度计测量每种颜色的温度。 温度从色谱的紫色到红色逐渐升高。然后，赫歇尔决定测量超出红色光的温度。尽管肉眼不可见，但该区域的温度读数最高。

Mandarin article C 有一回，北风跟太阳正在那儿争论谁的本领大。说着说着，来了一个路人，身上穿了一件厚袍子。他们俩就商量好了，说，谁能先让这个路人把他的袍子脱下来，就算他的本领大。北风卯足了劲儿，拼命的吹。可是，他吹的越厉害，那个人就把他的袍子裹得越紧。到末了，北风没辙了，只好就算了。一会儿，太阳出来一晒，那个人马上就把袍子脱了下来。 所以， 北风不得不承认，还是太阳比他的本领大。

Mandarin poem A. 《梦与诗》胡适都是平常经验，
都是平常影象，
偶然涌到梦中来，
变幻出多少新奇花样!
都是平常情感，
都是平常言语，
偶然碰着个诗人，
变幻出多少新奇诗句!
醉过才知酒浓，
爱过才知情重;
你不能做我的诗，
正如我不能做你的梦。

Mandarin poem B. 《我不知道风是在哪一个方向吹》徐志摩我不知道风
是在哪一个方向吹
我是在梦中，
甜美是梦里的光辉。
我不知道风
是在那一个方向吹
我是在梦中，
她的负心，我的伤悲。
我不知道风
是在哪一个方向吹
我是在梦中，
在梦的悲哀里心碎!

Mandarin prompts for spontaneous speech 是什么吸引你来到伦敦的?
你喜欢伦敦的哪些方面呢?
你不喜欢伦敦的哪些方面呢?
你有喜欢的餐馆或者菜系吗? 它在伦敦有吗?
你有去过任何伦敦的公园吗? 你最喜欢的是哪一个?
如果你碰到了一个刚来伦敦的人，你会建议他们去看什么或者做什么?
你有去过任何伦敦的景点吗 (比如，伦敦眼)? 你喜欢那个景点吗? 还是觉得它没有想象的那么好?

[1]See https://github.com/alexisdmacintyre/AcousticLandmarks (Last viewed July 9, 2021).

Abercrombie, D. (**1964**). "A phonetician's view of verse structure," Linguistics **2**(6), 5–13.
Adi, Y., Keshet, J., Cibelli, E., Gustafson, E., Clopper, C., and Goldrick, M. (**2016**). "Automatic measurement of vowel duration via structured prediction," J. Acoust. Soc. Am. **140**(6), 4517–4527.
Alexandrou, A. M., Saarinen, T., Kujala, J., and Salmelin, R. (**2020**). "Cortical entrainment: What we can learn from studying naturalistic speech perception," Lang. Cognit. Neurosci. **35**(6), 681–693.
Arvaniti, A. (**2009**). "Rhythm, timing and the timing of rhythm," Phonetica **66**(1-2), 46–63.
Arvaniti, A. (**2012**). "The usefulness of metrics in the quantification of speech rhythm," J. Phonetics **40**(3), 351–373.

Assaneo, M. F., Ripollés, P., Orpella, J., Lin, W. M., de Diego-Balaguer, R., and Poeppel, D. (**2019**). "Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning," Nat. Neurosci. **22**(4), 627–632.

Barbosa, P. A., Arantes, P., Meireles, A. R., and Vieira, J. M. (**2005**). "Abstractness in speech-metronome synchronisation: *P*-centres as cyclic attractors," in *Ninth European Conference on Speech Communication and Technology*.

Benveniste, E. (**1971**). "The notion of rhythm in its linguistic expression," in *Problems in General Linguistics* (University of Miami Press, Coral Gables, Florida), pp. 281–288.

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (**2017**). "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario," IEEE Trans. Neural Syst. Rehabil. Eng. **25**(5), 402–412.

Boersma, P., and Weenink, D. (**2020**). "Praat: Doing phonetics by computer (version 6.1.16) [computer program]" available at http://www.praat.org/ (Last viewed June 6, 2002).

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (**1992**). "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152.

Braiman, C., Fridman, E. A., Conte, M. M., Voss, H. U., Reichenbach, C. S., Reichenbach, T., and Schiff, N. D. (**2018**). "Cortical response to the natural speech envelope correlates with neuroimaging evidence of cognition in severe brain injury," Curr. Biol. **28**(23), 3833–3839.

Bramble, D. M., and Carrier, D. R. (**1983**). "Running and breathing in mammals," Science **219**(4582), 251–256.

Breska, A., and Ivry, R. B. (**2016**). "Taxonomies of timing: Where does the cerebellum fit in?," Curr. Opin. Behav. Sci. **8**, 282–288.

Bröhl, F., and Kayser, C. (**2021**). "Delta/theta band EEG differentially tracks low and high frequency speech-derived envelopes," Neuroimage **233**, 117958.

Caetano, M., and Rodet, X. (**2011**). "Improved estimation of the amplitude envelope of time-domain signals using true envelope cepstral smoothing," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York), pp. 4244–4247.

Chow, I., Belyk, M., Tran, V., and Brown, S. (**2015**). "Syllable synchronization and the *p*-center in cantonese," J. Phonetics **49**, 55–66.

Cummins, F. (**2012a**). "Looking for rhythm in speech," Empirical Musicol. Rev. **7**, 28.

Cummins, F. (**2012b**). "Oscillators and syllables: A cautionary note," Front. Psychol. **3**, 364.

Cummins, F. (**2014**). "The remarkable unremarkableness of joint speech," in *Proceedings of the 10th International Seminar on Speech Production*, ISSP, Cologne, DE, Vol. 73.

Cummins, F. (**2019**). *The Ground from Which We Speak: Joint Speech and the Collective Subject* (Cambridge Scholars Publishing, Newcastle upon Tyne, UK).

Dellwo, V., Leemann, A., and Kolly, M.-J. (**2015**). "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," J. Acoust. Soc. Am. **137**(3), 1513–1528.

Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (**2018**). "Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech," eNeuro. **5**(2), ENEURO.0084-18.2018.

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (**2016**). "Cortical tracking of hierarchical linguistic structures in connected speech," Nat. Neurosci. **19**(1), 158–164.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., and Poeppel, D. (**2017**). "Temporal modulations in speech and music," Neurosci. Biobehav. Rev. **81**, 181–187.

Ding, N., and Simon, J. Z. (**2014**). "Cortical entrainment to continuous speech: Functional roles and interpretations," Front. Hum. Neurosci. **8**, 311.

Doelling, K. B., and Assaneo, M. F. (**2021**). "Neural oscillations are a start toward understanding brain activity rather than the end," PLoS Biol. **19** (5), e3001234.

Drennan, D. P., and Lalor, E. C. (**2019**). "Cortical tracking of complex sound envelopes: Modeling the changes in response with intensity," eNeuro. **6**(3), ENEURO.0082-19.2019.

Duanmu, S. (**2001**). "Stress in Chinese," in *Chinese Phonology in Generative Grammar*, edited by D. B. Xu (Academic press, San Diego, CA), pp. 117–138.

Fraisse, P. (**1978**). "Time and rhythm perception," in *Handbook of Perception: Vol. 8. Perceptional Coding*, edited by E. Carterette and M. Friedman (Academic Press, New York), pp. 203–254.

Francis, A. L., Ciocca, V., and Ching Yu, J. M. (**2003**). "Accuracy and variability of acoustic measures of voicing onset," J. Acoust. Soc. Am. **113**(2), 1025–1032.

Franich, K. (**2018**). "Tonal and morphophonological effects on the location of perceptual centers (*p*-centers): Evidence from a Bantu language," J. Phonetics **67**, 21–33.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (**1993**). "Darpa TIMIT acoustic-phonetic continous speech corpus CD-ROM," NIST Speech Disc 1-1.1, NASA STI/Recon Technical Report No. 93, 27403.

Gervain, J., and Geffen, M. N. (**2019**). "Efficient neural coding in auditory and speech perception," Trends Neurosci. **42**(1), 56–65.

Ghitza, O. (**2011**). "Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm," Front. Psychol. **2**, 130.

Ghitza, O. (**2013**). "The theta-syllable: A unit of speech information defined by cortical function," Front. Psychol. **4**, 138.

Giraud, A.-L., and Poeppel, D. (**2012**). "Cortical oscillations and speech processing: Emerging computational principles and operations," Nat. Neurosci. **15**(4), 511–517.

Goldsmith, J. (**2011**). "The syllable," in *The Handbook of Phonological Theory*, edited by J. Goldsmith, J. Riggle, and A. Yu (Wiley-Blackwell, Oxford, UK), pp. 162–196.

Goslin, J., and Frauenfelder, U. H. (**2001**). "A comparison of theoretical and human syllabification," Lang. Speech **44**(4), 409–436.

Grabe, E., and Low, E. L. (**2002**). "Durational variability in speech and the rhythm class hypothesis," in *Papers in Laboratory Phonology*, edited by N. Warner and C. Gussenhoven (Mouton de Gruyter, Berlin), Vol. 7, pp. 515–546.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., and Garrod, S. (**2013**). "Speech rhythms and multiplexed oscillatory sensory coding in the human brain," PLoS Biol. **11**(12), e1001752.

He, L., and Dellwo, V. (**2016**). "A Praat-based algorithm to extract the amplitude envelope and temporal fine structure using the Hilbert transform," in *Proceedings of Interspeech 2016*, pp. 530–534.

Hoequist, C., Jr. (**1983**). "Syllable duration in stress-, syllable- and mora-timed languages," Phonetica **40**(3), 203–237.

Howell, P., and Kadi-Hanifi, K. (**1991**). "Comparison of prosodic properties between read and spontaneous speech material," Speech Commun. **10**(2), 163–169.

Jarne, C. (**2018**). "A heuristic approach to obtain signal envelope with a simple software implementation," Anales AFA **29**(2), 51–57.

Jones, M. R. (**1976**). "Time, our lost dimension: Toward a new theory of perception, attention, and memory," Psychol. Rev. **83**(5), 323–355.

Keitel, A., Gross, J., and Kayser, C. (**2018**). "Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features," PLoS Biol. **16**(3), e2004473.

Kohler, K. J. (**2009**). "Whither speech rhythm research?," Phonetica **66**, 5–14.

Kojima, K., Oganian, Y., Cai, C., Findlay, A., Chang, E., and Nagarajan, S. (**2021**). "Low-frequency neural tracking of speech amplitude envelope reflects the convolution of evoked responses to acoustic edges, not oscillatory entrainment," bioRxiv.

Kolly, M.-J., and Dellwo, V. (**2014**). "Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition," J. Phonetics **42**, 12–23.

Kumar, A., Shahnawazuddin, S., and Pradhan, G. (**2017**). "Non-local estimation of speech signal for vowel onset point detection in varied environments," in *INTERSPEECH*, pp. 429–433.

Lai, C., Sui, Y., and Yuan, J. (**2010**). "A corpus study of the prosody of polysyllabic words in Mandarin Chinese," in *Speech Prosody 2010—Fifth International Conference*.

Leow, L.-A., and Grahn, J. A. (**2014**). "Neural mechanisms of rhythm perception: Present findings and future directions," in *Neurobiology of Interval Timing*, edited by H. Merchant and V. de Lafuente (Springer, New York), pp. 325–338.

Lin, H., and Wang, Q. (**2007**). "Mandarin rhythm: An acoustic study," J. Chin. Lang. Comput. **17**(3), 127–140.

Marcus, S. M. (**1981**). "Acoustic determinants of perceptual center (*p*-center) location," Percept. Psychophys. **30**(3), 247–256.
Meyer, L., and Gumbert, M. (**2018**). "Synchronization of electrophysiological responses with speech benefits syntactic information processing," J. Cognit. Neurosci. **30**(8), 1066–1074.
Meyer, L., Henry, M. J., Gaston, P., Schmuck, N., and Friederici, A. D. (**2017**). "Linguistic bias modulates interpretation of speech via neural delta-band oscillations," Cerebral Cortex **27**(9), 4293–4302.
Meyer, L., Sun, Y., and Martin, A. E. (**2020**). "Synchronous, but not entrained: Exogenous and endogenous cortical rhythms of speech and language processing," Lang. Cognit. Neurosci. **35**(9), 1089–1099.
Molinaro, N., and Lizarazu, M. (**2018**). "Delta (but not theta)-band cortical entrainment involves speech-specific processing," Eur. J. Neurosci. **48**(7), 2642–2650.
Nolan, F., and Asu, E. L. (**2009**). "The pairwise variability index and coexisting rhythms in language," Phonetica **66**(1-2), 64–77.
Nolan, F., and Jeon, H.-S. (**2014**). "Speech rhythm: A metaphor?," Philos. Trans. R. Soc., B **369**(1658), 20130396.
Oganian, Y., and Chang, E. F. (**2019**). "A speech envelope landmark for syllable encoding in human superior temporal gyrus," Sci. Adv. **5**(11), eaay6279.
O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (**2015**). "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," Cerebral Cortex **25**(7), 1697–1706.
Patel, A. D., Löfqvist, A., and Naito, W. (**1999**). "The acoustics and kinematics of regularly timed speech: A database and method for the study of the *p*-center problem," in *Proceedings of the 14th International Congress of Phonetic Sciences*, Linguistics Department, University of California Berkeley, Vol. 1, pp. 405–408.
Peelle, J. E., and Davis, M. H. (**2012**). "Neural oscillations carry speech rhythm through to comprehension," Front. Psychol. **3**, 320.
Pefkou, M., Arnal, L. H., Fontolan, L., and Giraud, A.-L. (**2017**). "θ-band and β-band neural activity reflects independent syllable tracking and comprehension of time-compressed speech," J. Neurosci. **37**(33), 7930–7938.
Pope, M. A., and Studenka, B. E. (**2019**). "Experience with event timing does not alter emergent timing: Further evidence for robustness of event and emergent timing," J. Motor Behav. **51**(1), 113–120.
Presacco, A., Simon, J. Z., and Anderson, S. (**2016**). "Effect of informational content of noise on speech representation in the aging midbrain and cortex," J. Neurophysiol. **116**(5), 2356–2367.
Ramus, F., Nespor, M., and Mehler, J. (**1999**). "Correlates of linguistic rhythm in the speech signal," Cognition **73**(3), 265–292.
Räsänen, O., Doyle, G., and Frank, M. C. (**2018**). "Pre-linguistic segmentation of speech into syllable-like units," Cognition **171**, 130–150.
Rathcke, T., Lin, C.-Y., Falk, S., and Dalla Bella, S. (**2021**). "Tapping into linguistic rhythm," Lab. Phonology: J. Assoc. Lab. Phonol. **12**(1), 11.
Revathi, A., Sasikaladevi, N., Nagakrishnan, R., and Jeyalakshmi, C. (**2018**). "Robust emotion recognition from speech: Gamma tone features and models," Int. J. Speech Technol. **21**(3), 723–739.
Schachtenhaufen, R. (**2010**). "Looking for lost syllables in Danish spontaneous speech," in *Linguistic Theory and Raw Sound*, Copenhagen Studies in Language 40, edited by P. Juel Henrichsen (Samfundslitteratur, Frederiksberg), 61–88.
Schimmel, S., and Atlas, L. (**2005**). "Coherent envelope detection for modulation filtering of speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)* (IEEE, New York), Vol. 1, pp. I–221.
Schotola, T. (**1984**). "On the use of demisyllables in automatic word recognition," Speech Commun. **3**(1), 63–87.
Schuppler, B. (**2017**). "Rethinking classification results based on read speech, or: Why improvements do not always transfer to other speaking styles," Int. J. Speech Technol. **20**(3), 699–713.
Schuppler, B., Ernestus, M., Scharenborg, O., and Boves, L. (**2011**). "Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions," J. Phonetics **39**(1), 96–109.
Scott, S. K. (**1998**). "The point of *p*-centres," Psychol. Res. **61**(1), 4–11.
Seifart, F., Meyer, J., Grawunder, S., and Dentel, L. (**2018**). "Reducing language to rhythm: Amazonian Bora drummed language exploits speech rhythm for long-distance communication," R. Soc. Open Sci. **5**(4), 170354.
Shao, Y., and Wang, D. (**2008**). "Robust speaker identification using auditory features and computational auditory scene analysis," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, New York), pp. 1589–1592.
Small, C. (**1998**). *Musicking: The Meanings of Performing and Listening* (University Press of New England, Hanover, NH).
Strauß, A., and Schwartz, J.-L. (**2017**). "The syllable in the light of motor skills and neural oscillations," Lang. Cognit. Neurosci. **32**(5), 562–569.
Šturm, P., and Volín, J. (**2016**). "*P*-centres in natural disyllabic Czech words in a large-scale speech-metronome synchronization experiment," J. Phonetics **55**, 38–52.
Teki, S., Grube, M., and Griffiths, T. D. (**2012**). "A unified model of time perception accounts for duration-based and beat-based timing mechanisms," Front. Integr. Neurosci. **5**, 90.
Teki, S., Grube, M., Kumar, S., and Griffiths, T. D. (**2011**). "Distinct neural substrates of duration-based and beat-based auditory timing," J. Neurosci. **31**(10), 3805–3812.
Teoh, E. S., Cappelloni, M. S., and Lalor, E. C. (**2019**). "Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features," Eur. J. Neurosci. **50**(11), 3831–3842.
Tierney, A., and Kraus, N. (**2015**). "Evidence for multiple rhythmic skills," PLoS One **10**(9), e0136645.
Tilsen, S., and Arvaniti, A. (**2013**). "Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages," J. Acoust. Soc. Am. **134**(1), 628–639.
Valero, X., and Alias, F. (**2012**). "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," IEEE Trans. Multimedia **14**(6), 1684–1689.
Van Canneyt, J., Wouters, J., and Francart, T. (**2020**). "From modulated noise to natural speech: The effect of stimulus parameters on the envelope following response," Hear. Res. **393**, 107993.
Vander Ghinst, M., Bourguignon, M., Niesen, M., Wens, V., Hassid, S., Choufani, G., Jousmäki, V., Hari, R., Goldman, S., and De Tiège, X. (**2019**). "Cortical tracking of speech-in-noise develops from childhood to adulthood," J. Neurosci. **39**(15), 2938–2950.
Vicenik, C., and Sundara, M. (**2013**). "The role of intonation in language and dialect discrimination by adults," J. Phonetics **41**(5), 297–306.
Villing, R. (**2010**). "Hearing the moment: Measures and models of the perceptual centre," Ph.D. thesis, National University of Ireland Maynooth.
Vos, P. G., Mates, J., and van Kruysbergen, N. W. (**1995**). "The perceptual centre of a stimulus as the cue for synchronization to a metronome: Evidence from asynchronies," Q. J. Exp. Psychol. Sect. A **48**(4), 1024–1040.
Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., and Mattys, S. L. (**2010**). "How stable are acoustic metrics of contrastive speech rhythm?," J. Acoust. Soc. Am. **127**(3), 1559–1569.
Xie, Z., McLoughlin, I., Zhang, H., Song, Y., and Xiao, W. (**2016**). "A new variance-based approach for discriminative feature extraction in machine hearing classification using spectrogram features," Digital Signal Process. **54**, 119–128.
Yi, H. G., Leonard, M. K., and Chang, E. F. (**2019**). "The encoding of speech sounds in the superior temporal gyrus," Neuron **102**(6), 1096–1110.
Zec, D. (**2007**). "The syllable," in *The Cambridge Handbook of Phonology*, edited by P. De Lacy (Cambridge University Press, Cambridge, UK), pp. 161–194.
Zhao, X., and Wang, D. (**2013**). "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, New York), pp. 7204–7208.
Zoefel, B., Ten Oever, S., and Sack, A. T. (**2018**). "The involvement of endogenous neural oscillations in the processing of rhythmic input: More than a regular repetition of evoked neural responses," Front. Neurosci. **12**, 95.
Zwicker, E., and Terhardt, E. (**1980**). "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. Am. **68**(5), 1523–1525.
Zwicker, E., Terhardt, E., and Paulus, E. (**1979**). "Automatic speech recognition using psychoacoustic models," J. Acoust. Soc. Am. **65**(2), 487–498.