10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

How to hit that beat: Testing acoustic anchors of rhythmic movement with speech

Chia-Yuan Lin[1], *Tamara Rathcke*[1]

[1]English Language and Linguistics, University of Kent, UK

c.lin@kent.ac.uk, t.v.rathcke@kent.ac.uk

Abstract

Sensorimotor synchronisation with metronome and music have been extensively studied, while synchronisation with speech is still relatively poorly understood. The present study looks into the question how to define the best anchor of synchronised movement (finger tapping) in speech, and compares manually identified vowel onsets with four acoustic landmarks that were derived by different signal processing algorithms. Participants listened to repetitions of natural English sentences and were instructed to tap in synchrony with what they perceived to be the sentence beat. The time course of the sentences was tagged for a number of rhythmically relevant events, including vowel onsets, fastest energy increase (maxD), a combination of high local pitch and periodic energy (PPP), and the largest amplitude of intersyllabic and interstress timescales (IMF1 and IMF2). Vowel onsets and maxD showed consistent tapping patterns, while other landmarks performed worse than vowel onsets. These findings suggest that local energy changes shape sensorimotor synchronisation with speech and that energy contours might serve as anchors of rhythmic attention in spoken language.

Index Terms: speech rhythm, rhythm perception, sensorimotor synchronisation, automatic annotation, p-centre

1. Introduction

Auditory encoding of temporal information is known to be shaped and enhanced by motor activities [1]–[3]. When asked to hit the beat of a sound sequence, we can synchronise our movements to the beat we perceive and adjust for potential discrepancies through an auditory-motor feedback loop: We would accelerate our movements if we are behind the beat or we would slow down if we are ahead of it [1]. Previous studies showed that both adults and infants recognized a metrical structure as more familiar when it was congruent with their movements [2], [3]. If moving to a sound, listeners are better at extracting the beat, whereas holding still makes the beat extraction more difficult [1].

Sensorimotor synchronisation paradigms measure time-locking of a movement to an auditory signal, and have been widely used to investigate rhythm perception in music [4]–[7]. In such paradigms, participants are usually instructed to synchronise to a sound which could be a metronome or a music excerpt at variable tempi. Among these paradigms, finger-tapping is most frequently used to study beat perception and individual synchronisation skills [5]. In this paradigm, people's perception of rhythmic structure in the signal is described through temporal intervals between taps and the variability of such intervals. Moreover, the calculation of the time distance between a tap and the acoustic onset of a rhythmic event can tell us how accurate these taps were matched to an external rhythm [6], [7].

However, unlike rhythmic structure of Western music, rhythmic patterns of speech have been a matter of many debates [8]–[12]. If the finger tapping paradigm were to be used with speech, the target of synchronisation would have to be defined in the continuous speech signal [13]. Previous research debated the issue of the target of rhythmic attention in the speech signal, and put forward the idea of a 'perceptual centre', or 'p-centre' [14], [15]. A 'p-centre' of a speech stimulus is described as its 'psychological moment of occurrence' [14]. The 'p-centre' is said to be located around a vowel onset [16], [17], but it also has been shown to shift to the left/right of the acoustic vowel onset. Such shifts are influenced by the presence of preceding and following consonants, their type and duration [15], [16], [18], [19]. Other acoustic factors may also play a role, such as the duration of the vowel [20] and properties of its amplitude envelope [21]–[23], to name just a few. Moreover, these acoustic factors are likely to be interdependent and interact or compete with each other in some complex ways [15], [21], which means that a 'p-centre' location may be influenced by the properties of the whole syllable [16]. Attempts have been made to define the 'p-centre' with reference to the timing of articulatory gestures [24], though considerable individual differences seem to obscure a potential kinematic correlate of the 'p-centre' [25]. Overall, this line of research has so far not helped to solve the puzzle of the rhythmic target in speech since the 'p-centre' is difficult to define and lacks the necessary robust cues [26].

Recently, we showed that in simple speech stimuli, the best predictor of a tap location was the vowel onset [13]. Participants' task was to synchronise with syllables (/bi/, /bu/) or tones (High, Low). The results showed that tapping asynchronies found with the acoustic onsets of tonal stimuli were comparable in magnitude and statistically identical to asynchronies found with the vowel onsets in verbal stimuli. Two issues arise from this research: Firstly, vowel onsets need to be identified manually while manual segmentation is sometimes difficult, especially when a vowel is preceded by a sonorant [27]. Moreover, vowel onsets also coincide with the onset of voicing and pitch, which might explain why they are readily available as the tap attractors. For example, co-speech gestures tend to coincide with local f0 maxima [28]. Similarly, a high-pitched syllable is likely to be interpreted as stressed, even if temporarily [29].

The aim of the present study is to evaluate several automated procedures that describe acoustic properties of speech signals, and investigate their ability to automatically predict tap location, in comparison to vowel onsets. Three algorithms were chosen:

10.21437/SpeechProsody.2020-1

- *Maximal energy increase - maxD* [16]: maxD has been discussed as the closest approximate of a 'p-centre' in Czech [16], and is calculated as the moment of the fastest energy increase in the amplitude envelope around a syllable nucleus.
- *Periodic Power Praat - PPP* [30]–[32]: PPP describes acoustic properties of F0 by taking into account concurrent signal energy, thus potentially helping to distinguish between vowels and sonorants and to identify perceptually salient moments of pitch contours.
- *Empirical Mode Decomposition - EMD* [33]: EMD identifies intrinsic mode functions (IMF) of speech signals, which have been shown to represent different time-scales of speech. The first and the second IMF frequencies seem to represent syllable and stress rates, and are known to be significantly different across languages [33].

The three algorithms differ in their complexity and the number of acoustic dimensions they take into account. While maxD relies exclusively on an energy derivate of the amplitude envelope, PPP integrates energy and F0. In contrast, EMD is derived from energy, frequency and time of speech spectra and is the most comprehensive representation of potentially relevant rhythmic properties of speech.

The present study will compare the three algorithms to the vowel onset that was annotated manually, with the aim to empirically evaluate which of the methods can best describe finger tapping performance with natural speech. The following two measures of tapping performance will be examined: (1) the number of taps produced in the proximity of an identified acoustic landmark, and (2) the local proximity between a temporal landmark and a tap. The higher the number of taps produced near a landmark, the higher the accuracy, the better a tap attractor the corresponding acoustic landmark would be.

2. Method

2.1. Stimuli

Six English sentences recorded by a young female speaker, native of the Standard Southern British English. They were chosen from an existing database ([34], see Table 1). The sentences were repeated 20 times, with a 400 ms pause between each repetition.

Table 1: *Summaries of the sentence materials.*

Sentence length	Sentence	Duration (secs)
Short: 4 syllables	S1: *I wove a yarn.*	1.1
	S2: *I took the prize.*	1.2
Medium: 7 syllables	M1: *Ann Won the yellow award*	1.5
	M2: *Grandpa did not eat the cake*	1.8
Long: 10 syllables	L1: *The incident occurred last Friday night*	2.1
	L2: *As the boy sneezed, the door closed suddenly.*	3.0

2.2. Acoustic landmarks

Syllable boundaries and vowel onsets were manually annotated by a trained phonetician (the second author). Subsequently, following local temporal landmarks were identified for each pre-defined syllable:

- *maxD – landmark of the fastest energy change*. Smoothed energy contours were firstly created following the procedure developed by Šturm and Volín [16]. A maxD value was derived from the difference energy contour for each syllable.
- *PPP – landmark of the concurrent high frequency and energy*. The intensity and F0 contour were derived from a Praat plug-in and R combined script [31], [32]. The moment of the highest F0 and periodic power was extracted for each syllable by picking out the largest product of these two values.
- *EMD – landmarks of the highest instantaneous frequencies of IMF1 and IMF2*. The procedure to generate the IMF components followed Tilsen and Arvaniti's work [33]. The moment of local maximum amplitude of the first and the second IMF was extracted, resulting in two landmarks. The local maximum amplitudes smaller than 25% of the largest maximum amplitude were excluded.

Figure 1 shows an example of the word "prize" (S2) and compares the timepoint of the vowel onset with the timepoints of the four acoustic landmarks derived by the three algorithms. As can be seen, locations can be variable. IMF1 and IMF2 have identical timepoints in stressed (as in "prize") but not in unstressed syllables.

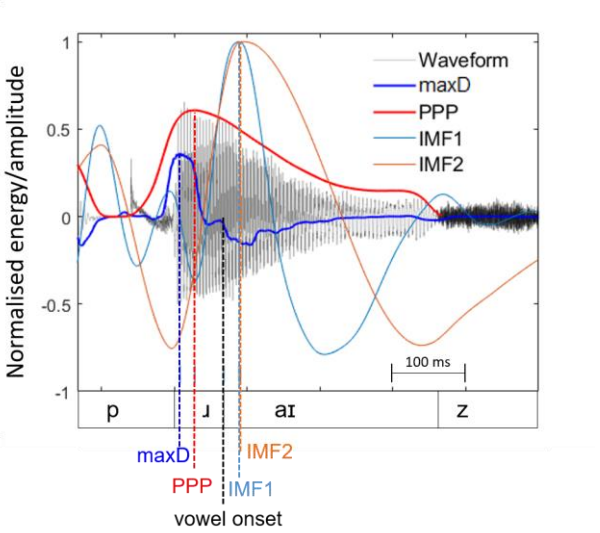

Figure 1. *Waveform and annotation of the example word "prize" (taken from S2), comparing locations of vowel onsets, maxD, PPP, IMF1 and IMF2.*

2.3. Participants

Twenty-nine native English speakers participated the experiment (21 female; mean age 23.1 years, range 18 – 36 years). All participants had no language impairment, motor disorders, or hearing problems. Participants gave an informed consent and received a small fee in compensation for their time. No professional musicians or dancers were among participants. Fifteen participants had taken amateur music lessons (mean experience among them: 4.3 years). Twelve had taken dancing lessons and four of them self-reported as an experienced dancer.

2.4. Task, procedure, and apparatus

Participants had to listen to looped stimuli and were instructed to start tapping with the finger of their dominant hand in

synchrony with the perceived beat of the stimulus as soon as they were able to. The sentences were given in a fixed order, from the shortest to the longest one (S1-L2, see Table 1). Tapping responses were recorded in the MIDI format using a Roland HPD-20 percussion pad and a Dell Latitude 7390 laptop. The CakeWalk by BandLab freeware controlled the play of spoken sentences and recorded the tapping responses.

2.5. Pre-processing of the tapping data

A tapping distribution was constructed for each participant by applying a Gaussian kernel density estimation to all taps obtained for each sentence. Figure 2 shows an example of the group densities for the test sentence S2 ("*I took the prize*").

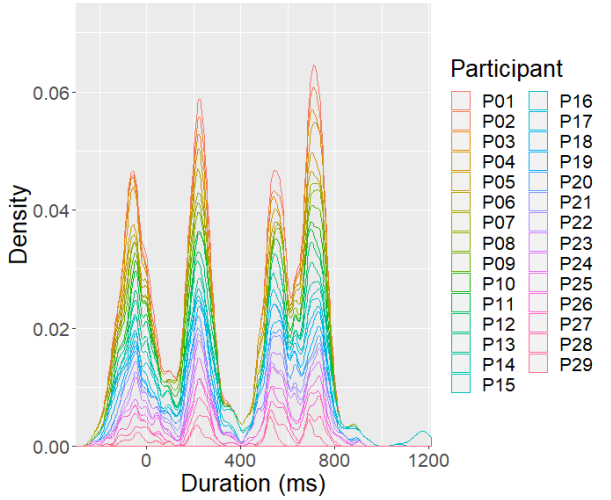

Figure 2. *Tapping probabilities for the sentence S2, aggregated across all participants' data.*

2.6. Analyses

To answer research questions of the present study, the following measures of tapping performance were calculated:

- *Normalised tap frequency*: The number of taps produced in the proximity of the acoustic landmarks (±120 ms) was calculated for each individual and sentence, and then normalised with the reference to the total number of taps produced for the sentence by an individual.
- *Asynchronies*: The tapping peaks of the aggregated tapping distribution (Figure 2) were used. Only the peaks over 40% of the maximum peak were included. Absolute asynchronies between the tapping peaks and the closest landmarks (±120 ms) were then calculated for each individual and sentence.

A series of linear mixed-effect models were fit to these measures as dependent variables. The fixed predictor was the acoustic landmark, running four planned comparisons: vowel onset vs. maxD, vowel onset vs. PPP, vowel onset vs. IMF1, and vowel onset vs. IMF2. Three random intercepts were included in the models: participant, sentence and the serial order of the landmark in a sentence (unless there were model convergence issues that led to the removal of one random effect).

3. Results

3.1. Normalised tap frequency

Planned comparisons showed that more taps were produced around vowel onsets than PPP (t=-2.31, p=.02), IMF1 (t=-2.73, p=.006), and IMF2 (t=-9.90, p<.001) while the number of taps around maxD was no different from the number of taps recorded around vowel onsets (t=0.57, p=.57). However, under a Bonferroni-corrected p = 0.0125 (α=.05/4) that we applied after running four comparisons, the difference between vowel onsets and PPP was only marginally significant. Figure 3 shows model estimations of this normalized tapping frequency measure.

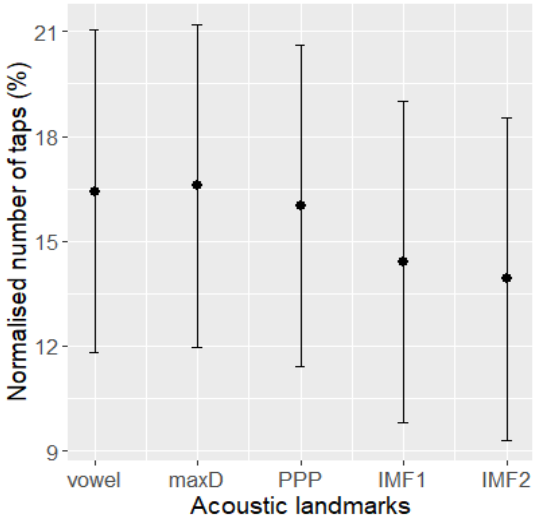

Figure 3. *Percentage of taps produced in the proximity of the acoustic landmarks.*

3.2. Asynchronies

Planned comparisons showed that absolute asynchronies were smaller for vowel onsets than for PPP (t=5.92, p<.001), IMF1 (t=12.68, p<.001) and IMF2 (t=14.34, p<.001) while the slight numerical difference between vowel onsets and maxD was not significant (t=0.60, p=.55). Figure 4 displays estimated magnitudes of the effects, plotted from the fitted mixed-effects models.

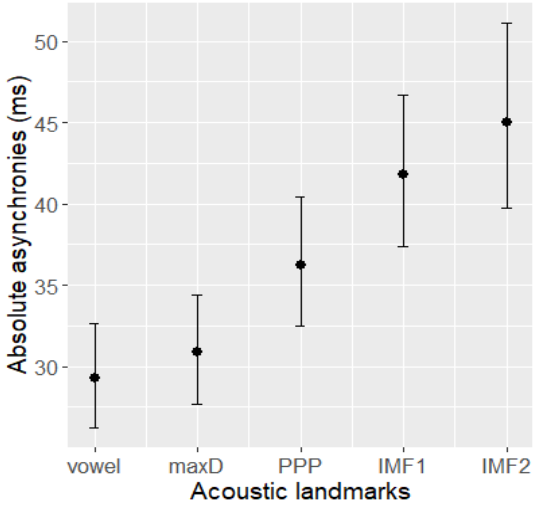

Figure 4. *Absolute asynchronies between acoustic landmarks and taps.*

4. Discussion

The present study aimed to identify the best acoustic anchor of sensorimotor synchronisation in speech. We compared four acoustically derived landmarks to manually identified vowel onsets. The acoustic landmarks included the fastest energy increase (maxD), a combination of F0 frequency and periodic energy (PPP) and the largest amplitude on the timescale of intersyllabic and interstress intervals (IMF1 and IMF2). The results showed that maxD performed no different from the vowel onset in both attracting taps (as measured by the normalized tap frequencies) and capturing sensorimotor precision (as measured by absolute asynchronies). In contrast, PPP attracted a similar number of taps as the vowel onset did but had larger absolute asynchronies. The two IMF landmarks performed worse than the vowel onset on both measures.

We chose IMF1 and IMF2 as they have been shown to represent the intersyllabic and interstress timescales of natural speech which are language-specific [33]. The core strength of IMF1 and IMF2 for the present study stems from their purely signal-driven identification, i.e. they were the only acoustic landmarks that were generated without any reference to the pre-defined syllable boundaries. However, in terms of serving as consistent targets of sensorimotor synchronisation with speech, both IMFs performed significantly worse than the vowel onset as fewer taps fell around the local amplitude peaks of IMF1 and IMF2, and the taps were generally located far away from these amplitude peaks. Reasons for this finding might be two-fold. First, the algorithm excels at tracking a stable phase of syllables and stresses, but lacks temporal precision that is needed to represent the location of rhythmic events. Second, local IMF maxima that were considered as potential SMS anchors might be misrepresenting the timescales identified by IMF. For example, our recent work showed that local amplitude maxima of the amplitude envelope served as poor predictors of SMS location [13]. Instead, local IMF minima might be more suitable as indicators of the onsets of rhythmic events. This alternative will be tested in future work. Moreover, since our application of the IMF algorithms did not involve any manual annotation, the number of IMF1 and IMF2 showed discrepancies to the total number of syllables and stresses in a sentence. Such discrepancies need to be systematically investigated and tested if, and how, they influence the performance of IMFs as potential tap attractors.

The PPP landmarks rely on F0 height and periodic power simultaneously. Given that in spontaneous speech the apex of co-speech gestures often coincides with high pitch [28], [35], we expected PPP to serve as a good attractor of finger taps in a laboratory task like our SMS paradigm. However, the distance between tapping peaks and PPPs was not as close as the distance between tapping peaks and vowel onsets. Again, this finding might arise because local maxima of amplitude/energy do not seem to be attracting movement in synchrony with speech [13]. In the present study, we used the largest product of F0 frequency and periodic power in each syllable. Yet this combination can be amended as different weights can be given to the two components and their values. Perhaps high amplitude received too high a weighting in the current calculation of PPP. This possibility will be examined in future work.

Although PPP did not show low absolute asynchronies, it captured a substantial number of produced taps. Since the PPP algorithm offers a continuous F0 tracking while considering periodic energy simultaneously [32], it offers further avenues for future work. For example, the present study concentrated exclusively on high local F0 values, though some pitch accents in English can have a low pitch-accent tone [36]. In future work, we may hypothesise that sensorimotor synchronisation might be sensitive to the phonological composition of the continuous pitch contour of a sentence, and that PPP might help to capture it with a higher level of precision than any other algorithms in question. It is further worth noting that the PPP algorithm can also be used to locate local minima of the periodic energy curve, and that these time points may serve as a proxy for syllable onsets [32]. The PPP landmarks can thus be generated automatically (similar to the IMF landmarks of this study), increasing the strengths of this measure.

Finally, maxD performed as well as vowel onsets in terms of capturing tap frequency and accuracy. Previous work has shown that in cyclical speech production with a metronome, maxD was located close to the periodic metronome pulse [16]. The measure seems to isolate a relevant feature of rhythmic properties in repetitive speech. The location of maxD is argued to signify the location of ‘p-centre’ in Czech [16] and potentially other languages [14] , i.e. between the minimum and the maximum of a local amplitude envelope, slightly closer to its minimum rather than the maximum [37]. While there is still no reliable ‘p-centre’ algorithm to date [38], maxD offers a potential solution in Czech production data [16], and English tapping data of the present study. The only disadvantage of maxD is its reliance on pre-defined syllable onsets in order to compute. A combination of PPP and maxD or IMF and maxD might benefit maxD in making it run without any manual input.

The present study compared different acoustic landmarks in their ability to capture targets of rhythmic attention in speech that was accessed with an SMS task. The results indicate that neither local high pitch nor by local energy maxima arising at an intersyllabic or interstressed rate were well suited to predict the temporal location of finger taps. Instead, the target of rhythmic attention in speech appears to be best captured by the fastest energy increase. That is, the energy increase and resulting spectral dysfluencies might be the key to the dynamics of the auditory-motor feedback loop when synchronising with spoken sentences.

5. Conclusions

The present study investigated the anchor of sensorimotor synchronisation in repetitive spoken sentences by comparing manually identified vowel onsets with automatically/semi-automatically generated acoustic landmarks. The moment of local maximal energy increase (maxD) performed as well as vowel onsets did, while other acoustic landmarks were less well suited to capture SMS performance in the data. These findings indicate that rhythmic attention in language might lock on to energy contours and their fluctuations [39] rather than local amplitude maxima [13]or pitch cues. However, the results of the present study are currently limited to SMS with only six English sentences, and require replication with a larger set of sentences. The present conclusions should therefore be considered preliminary.

6. Acknowledgements

The authors would like to thank Aviad Albert and Francesco Cangemi for their help with the PPP algorithm, Sam Tilsen for his support with the IMF analyses, and Pavel Šturm for sharing his maxD script. Our special thanks go to Georg Lohfink for his assistance with the data collection.

7. References

[1] Y.-H. Su and E. Pöppel, “Body movement enhances the extraction of temporal structures in auditory sequences,” *Psychological Research*, vol. 76, no. 3, pp. 373–382, 2012.

[2] J. Phillips-Silver and L. J. Trainor, “Psychology: Feeling the beat: Movement influences infant rhythm perception,” *Science*, vol. 308, no. 5727, p. 1430, 2005.

[3] J. Phillips-Silver and L. J. Trainor, “Hearing what the body feels: Auditory encoding of rhythmic movement,” *Cognition*, vol. 105, no. 3, pp. 533–546, 2007.

[4] G. Aschersleben, “Temporal control of movements in sensorimotor synchronization,” *Brain and Cognition*, vol. 48, no. 1, pp. 66–79, 2002.

[5] S. Dalla Bella, N. Farrugia, C. E. Benoit, V. Begel, L. Verga, E. Harding, and S. A. Kotz, “BAASTA: Battery for the Assessment of Auditory Sensorimotor and Timing Abilities,” *Behavior Research Methods*, vol. 49, no. 3, pp. 1128–1145, Jun. 2017.

[6] B. H. Repp, “Sensorimotor synchronization: A review of the tapping literature,” *Psychonomic Bulletin and Review*, vol. 12, no. 6, pp. 969–992, 2005.

[7] B. H. Repp and Y.-H. Su, “Sensorimotor synchronization: A review of recent research (2006-2012),” *Psychonomic Bulletin and Review*, vol. 20, no. 3, pp. 403–452, 2013.

[8] S. Dalla Bella, A. Białuńska, and J. Sowiński, “Why Movement Is Captured by Music, but Less by Speech: Role of Temporal Regularity,” *PLoS ONE*, vol. 8, no. 8, pp. 1–16, 2013.

[9] R. M. M. Dauer, “Stress-timing and syllable-timing reanalyzed,” *Journal of Phonetics*, vol. 11, no. 1, pp. 51–62, 1983.

[10] A. Arvaniti, “Rhythm, timing and the timing of rhythm,” *Phonetica*, vol. 66, no. 1–2, pp. 46–63, 2009.

[11] A. Arvaniti, “The usefulness of metrics in the quantification of speech rhythm,” *Journal of Phonetics*, vol. 40, no. 3, pp. 351–373, May 2012.

[12] A. D. Patel and J. R. Daniele, “An empirical comparison of rhythm in language and music.,” *Cognition*, vol. 87, no. 1, pp. B35-45, Feb. 2003.

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

[14] J. Morton, S. M. Marcus, and C. Frankish, “Perceptual centers (P-centers),” *Psychological Review*, vol. 83, no. 5, pp. 405–408, 1976.

[15] S. M. Marcus, “Acoustic determinants of perceptual center (P-center) location,” *Perception & Psychophysics*, vol. 30, no. 3, pp. 247–256, 1981.

[16] P. Šturm and J. Volín, “P-centres in natural disyllabic Czech words in a large-scale speech-metronome synchronization experiment,” *Journal of Phonetics*, vol. 55, pp. 38–52, 2016.

[17] P. A. Barbosa, P. Arantes, A. R. Meireles, and J. M. Vieira, “Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors,” in *9th European Conference on Speech Communication and Technology*, 2005, pp. 1441–1444.

[18] G. D. Allen, “The location of rhythmic stress beats in English : an Experimental Study I,” *Language and Speech*, vol. 15, no. 1, pp. 72–100, 1972.

[19] R. A. Fox and I. Lehiste, “The effect of final consonant structure on syllable onset location,” *The Journal of the Acoustical Society of America*, vol. 77, no. S1, pp. S54–S54, Apr. 1985.

[20] R. A. Fox and I. Lehiste, “The effect of vowel quality variations on stress-beat location,” *Journal of Phonetics*, vol. 15, pp. 1–13, Jan. 1987.

[21] B. Pompino-Marschall, “On the psychoacoustic nature of the P-center phenomenon,” *Journal of Phonetics*, vol. 17, no. 3, pp. 175–192, Jul. 1989.

[22] C. A. Harsin, “Perceptual-center modeling is affected by including acoustic rate-of-change modulations,” *Perception and Psychophysics*, vol. 59, no. 2, pp. 243–251, 1997.

[23] P. Howell, “Prediction of P-center location from the distribution of energy in the amplitude envelope: I & II,” *Perception & Psychophysics*, vol. 43, no. 1, p. 99, 1988.

[24] C. A. Fowler, “‘Perceptual centers’ in speech production and perception,” *Perception & Psychophysics*, vol. 25, no. 5, pp. 375–388, 1979.

[25] K. J. De Jong, “The correlation of P-center adjustments with articulatory and acoustic events,” *Perception & Psychophysics*, vol. 56, no. 4, pp. 447–460, Jul. 1994.

[26] R. C. Villing, B. H. Repp, T. E. Ward, and J. M. Timoney, “Measuring perceptual centers using the phase correction response,” *Attention, Perception, and Psychophysics*, vol. 73, no. 5, pp. 1614–1629, Jul. 2011.

[27] A. Turk, S. Nakai, and M. Sugahara, “Acoustic Segment Durations in Prosodic Research: A Practical Guide,” in *Methods in Empirical Prosody Research*, 2012, pp. 1–28.

[28] N. Esteve-Gibert and P. Prieto, “Prosodic structure shapes the temporal realization of intonation and manual gesture movements,” *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 3, pp. 850–864, Jun. 2013.

[29] K. Zahner, S. Kutscheid, and B. Braun, “Alignment of f0 peak in different pitch accent types affects perception of metrical stress,” *Journal of Phonetics*, vol. 74, pp. 75–95, 2019.

[30] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh, “Use of temporal information: Detection of periodicity, aperiodicity, and pitch in speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 776–786, Sep. 2005.

[31] A. Albert, F. Cangemi, and M. Grice, “Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration,” in *Proc. 9th International Conference on Speech Prosody 2018*, 2018, pp. 804–808.

[32] F. Cangemi, A. Albert, and M. Grice, “Modelling intonation: Beyond segments and tonal targets,” in *Proceedings of the 19th International Congress of Phonetic Sciences*, 2019, pp. 572–576.

[33] S. Tilsen and A. Arvaniti, “Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages,” *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 628–639, Jul. 2013.

[34] T. V. Rathcke, S. Falk, and S. Dalla Bella, “Linguistic structure and listener characteristics modulate the ‘speech-to-song illusion,’” in *15th International Conference on Music Perception and Cognition*, 2018.

[35] W. Pouw, S. J. Harrison, and J. A. Dixon, “Gesture-Speech Physics: The Biomechanical Basis for the Emergence of Gesture-Speech Synchrony,” *Journal of Experimental Psychology: General*, 2019.

[36] D. Ladd, *Intonational phonology*. Cambridge, UK: Cambridge University Press, 2008.

[37] F. Cummins and R. Port, “Rhythmic constraints on stress timing in English,” *Journal of Phonetics*, vol. 26, no. 2, pp. 145–171, 1998.

[38] R. Villing, T. Ward, and J. Timoney, “P-Centre Extraction from Speech : the need for a more reliable measure,” in *Irish Signals and Systems Conference 2003*, 2003, pp. 5–10.

[39] U. Goswami and V. Leong, “Speech rhythm and temporal structure: converging perspectives,” *Laboratory Phonology*, vol. 4, no. 1, pp. 67–92, 2013.