Cognitive Constraints and Experience Mold Speech Rhythm: Evidence from Thai Speech Cycling

Francesco Burroni[1,2], Komtham Domrongchareon[3]
[1]Spoken Language Processing Group, Institute for Phonetics and Speech Processing, LMU München, Germany
[2]Center of Excellence in southeast Asian Linguistics, Faculty of Arts, Chulalongkorn University
[3]Faculty of Music, Silpakorn University, Bangkok, Thailand
francesco.burroni@phonetik.uni-muenchen.de
domrongchareon_k@silpakorn.edu

Abstract

This study explores phonological rhythm in Thai through the speech cycling (SC) paradigm. Six native Thai speakers, with and without musical training, produced phrases synchronized to external rhythmic cues. We measured the alignment of stressed syllables within a phrase repetition cycle (PRC) and analyzed the distribution of these alignments. The results revealed that Thai speakers consistently aligned stressed syllables at specific ratios, such as 1/3, 1/2, and 2/3 of the PRC. The study also found differences based on musical training, with trained participants showing more refined rhythmic patterns, suggesting a complex interplay of both universal and experience-based rhythmic constraints.

1 Introduction

The study of speech rhythm is a primary area of investigation in both the phonological and phonetic literature where the question has been approached from a variety of angles (*cf.* Turk & Shattuck-Hufnagel, 2013 for an in-depth review). Speech rhythm has predominantly been examined through the lens of cross-linguistic comparisons of so-called speech rhythm "metrics" in search of different rhythmic classes across languages. This approach that has yielded varied results, as will be discussed in detail in the next section (*cf.* Arvaniti, 2012; Bertinetto, 1989).

However, alternative approaches to the study of speech rhythm have been developed. Of specific interest here are attempts at grounding speech rhythm in more general cognitive constraints on speech production and perception (*cf.* Cummins & Port, 1998; Franich, 2021; Port, 2003; Tilsen, 2009).

In this paper, we follow this second family of approaches and conduct an experimental investigation of phonological rhythm in Thai using the "speech cycling" paradigm, an experimental task where participants have to produce words at specific points, known as phases, of a larger phrase cycle in accordance with an external rhythmic cue (Cummins & Port, 1998; Tajima & Port, 2003).

In the remainder of this introduction, we first introduce previous research on speech rhythm based on the rhythm class hypothesis and the challenges this approach has encountered. Subsequently, we introduce the speech cycling paradigm as way to overcome some of these limitations and to ground speech rhythm in more general cognitive mechanisms. Finally, we outline the suitability of Thai rhythm as a good case study given the dearth of experimental work on the topic and previous conflicting findings.

1.1 The rhythm class hypothesis

The notion that languages belong to different rhythmic classes, based on isochrony at the syllable ("syllable timing") or stress-interval levels ("stress timing"), was influentially proposed by Pike (1945) and Abercrombie (1990). However, experimental work probing isochrony failed to observe it (Arvaniti, 2009; Bertinetto, 1985; Dauer, 1983; Fletcher, 2010). The view of rhythm as isochrony was abandoned and a new notion of rhythm based on a complex interplay of language-specific phonological and syntactic properties emerged (Bertinetto, 1989; Dauer, 1983, 1987; Fletcher, 2010).

In the 1990s, following the work of Ramus and colleagues (Ramus et al., 1999), a renewed interest towards metrics that could help establishing rhythmic classes arose, e.g., (Bertinetto & Bertini, 2010; Dellwo, 2006; Grabe & Low, 2002). Such metrics mostly measure durations and include the proportion of vocalic intervals (%V), the standard deviation of consonantal and vocalic intervals (ΔC, ΔV) and their variation coefficients (varcoΔC, varcoΔV). Other metrics also include vocalic and intervocalic raw and normalized pairwise variability indices *(nPVI*, *rPVI*). Despite this renewed interest in rhythmic classes, a variety of problems with the proposed rhythmic metrics emerged (Arvaniti, 2009; Kohler, 2009).

First, the classification of languages with "unknown" or "mixed" rhythmic typologies turned out to be far from straightforward. For instance, Thai was classified as stress-timed using PVIs, but as syllable-timed using %V and ΔC (Grabe & Low, 2002). Second, it was also pointed out that the new metrics were highly sensitive to segmental materials (Arvaniti, 2009; Fletcher, 2010; Mairano & Romano, 2011). Third, a large crosslinguistic study demonstrated that rhythmic differences across languages – attributable to rhythmic classes – and confounds – like elicitation task and segmental composition – have comparable effects on rhythm metrics (Arvaniti, 2012).

Due to the challenges of studying speech rhythm using rhythm metrics applied to elicited or natural speech, several scholars have recommended a shift in focus. Instead of concentrating solely on timing properties, as captured by traditional rhythm metrics, they suggest examining higher-level patterns in grouping and prominence both in speech production and in listeners' perception (Arvaniti, 2009; Kohler, 2009).

An experimental paradigm, called "speech cycling" (Chung & Arvaniti, 2013; Cummins & Port, 1998; Franich, 2021; Tajima & Port, 2003; Tilsen, 2009; Zawaydeh et al., 2002), has been developed exactly as a mean to uncover constraints on prominence and grouping patterns and their relationship to speech rhythm.

1.2 The speech cycling paradigm

The "speech cycling" paradigm – henceforth SC – was first developed by Cummins and Port (1998). SC is a rhythmic task where participants produce words at specific points, known as phases, of a larger phrase cycle; they do so by entraining to an external rhythmic cue.

SC involves entraining the initial and final words of a short phrase – for example "*beg for a dime*" – to high (H) and low (L) metronome beats. While the H-L interval duration remains constant across trials, the duration of two successive H–H, called the phrase repetition cycle (PRC), is systematically manipulated. The final stressed syllable thus needs to be aligned at different phase – e.g., 0.3, 0.5, and 0.75 – of the PRC, Figure 1.

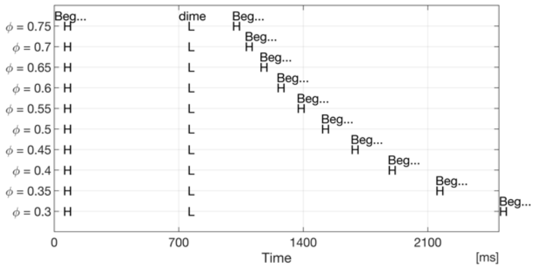

Figure 1: Illustration of word entrainment to H and L tones for different phases and duration of the PRC for the phrase "*beg for a dime*"

Thus, in SC, participants are exposed to a uniform rhythmic continuum of possible phases for the final stressed word within the larger phrases cycle. Thus, SC experiments can be used to probe whether a rhythmic continuum can be faithfully reproduced by participants or whether the continuum is warped into a small number of discrete categories, Figure 2.

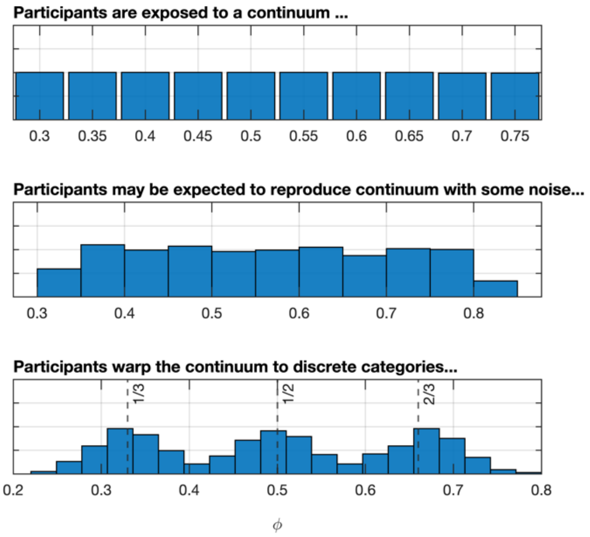

Figure 2: Logic of Speech Cycling experiment, see text for more details.

Cummins and Port (1998) found that American English (AE) speakers have a strong tendency to warp the rhythmic continuum intro three categories. Specifically, they produce stressed syllable at the harmonic series or its multiples (1/3, 1/2, 2/3) of the PRC, Figure 2. Cummins and Port (1998) demonstrated that these values of the PRC also exhibit lowest variance and can, thus, conceptualized as "attractors" in the potential landscape of a dynamical system. Thus, Cummins and Port (1998) and Port (2003) related the rhythmic warping they observed to the relative initiation of a new foot relative to PRC and conceptualized the process as a system of coupled oscillators. From this perspective, the different attractors they observed can also be translated into low-dimensional phonological representations. Specifically, the attractors were taken to reflect the initiation of a new trochaic metrical foot in AE.

An attractor at 1/3 is taken to reflect the metrical grouping [*beg for a*][*dime*][-], the only possible grouping where stress on the final syllables appears at 1/3 of the PRC. Similarly, the attractor an ½ represents the grouping [beg for a][dime]; and an attractor at 2/3 represents a grouping [beg][for a][dime]. These patterns exhaust the rhythmic possibilities of English speakers and reveal the organization of prominence and grouping in this language. SC allows us to understand that the impression of AE rhythm being driven by stressed syllables may arise in part from phonological properties (e.g., vowel reduction etc.) but, crucially, also from the strong constraints that exist on the distributions of stressed syllables within phrases. If the distribution is highly constrained, repetition of similar pattern will naturally arise resulting in "rhythmic" patterns.

A final important discovery of Cummins and Port (1998) was that the rhythmic warping observed in SC is identical for both rhythmically naïve participants, as well as rhythmically trained participants, e.g., professional musicians. These finding suggest that the constraints observed in speech cycling reside above experience, possibly being cognitive in nature.

Since the original study on AE, SC has been applied to other languages, to show that constraints exists on the production foot-initial syllables in Japanese like (Tajima & Port, 2003); stressed syllables in Arabic (Zawaydeh et al., 2002); accentual-phrase initial syllables in Korean (Chung & Arvaniti, 2013); and foot-initial syllable in Medʉmba (Franich, 2021). More work on AE (Tilsen, 2009) has also tried to further develop the conceptualization of rhythm observed in SC as a system of coupled oscillators by taking into account the initiation of articulatory gestures and their variability.

Despite the interest attracted by this paradigm, many aspects of SC remain underexplored. No work has investigated further the role of rhythmical/musical training on speakers' behavior during SC. Additionally, the number of languages investigated with speech cycling remains scarce. For instance, no Asian tonal language or prominence-final, so–called iambic, language has been investigated. Modelling work using dynamical systems outside of English is also lacking. With these issues in mind, we introduce the case study to which we applied the SC paradigm.

1.3 The case study: Thai Rhythm

There are several aspects that make Thai a good case study to investigate rhythm using SC.

First, the rhythmic class of Thai is debated. Thai has been described as both syllable timed (Pantupong, 1973; Suntornsawet, 2022) and stress-timed/mixed (Luangthongkum, 1978). Rhythm metrics have not settled the matter. The rhythmic classification depends on the metric used (Grabe & Low, 2002).

Second, if Thai really is as stress timed as some report (Mairano & Romano, 2011), it is quite different from AE in view of its tonal nature, simple phonotactics and, above all, iambic rhythm. In iambic rhythms, the nature of the foot type is driven by durational cues, naturally forming groupings with longer final prominent elements. In line with the iambic nature of Thai, duration has been often reported to be the primary cue to stress in Thai (Nitisaroj, 2004; Potisuk et al., 1996). This is opposed to trochaic systems where grouping is more intensity based (Hayes, 1995). Thus, probing the behavior of Thai speakers compared to AE speakers is of great interest.

Third, while prominence is uncontroversially final in Thai (Bee, 1975; Bennett, 1994), the grouping around prominent syllables is debated. Some assume cretic structures [–◡–] (Bee, 1975), while others have shown experimental evidence for anapests [◡◡–] (Gandour et al., 1992). Fourth, Thai rhythm has been hypothesize to display a high degree of individual variation (Luangthongkum, 1978).

1.4 Research Questions and Predictions

In view of the issues outlined in the previous sections, we focus on three research questions concerning Thai rhythm probed using the SC paradigm. These are:

1) Do Thai speakers exhibit rhythmic constraints in their production of stressed syllables within phrases, similar to AE speakers?

2) Taking individual music experience into account, are there differences in these behaviors based on rhythmic/musical training?

3) If rhythmic constraints are manifested, does this behavior betray the signatures of an underlying dynamical system?

We put forth the following predictions. For 1), given the warping of rhythmic continua in various languages, we expect to observe it in Thai too. However, in view of iambic nature of Thai, we also expect final prominence and different grouping compared to trochaic languages, like AE.

For 2), we expect, based on previous work on AE (Cummins & Port, 1998), a similar behavior for participants regardless of their musical/rhythmic background. For 3), under the assumption that speech cycling rhythm can be understood in terms of attractors in a dynamical system of coupled oscillators, we expect the attractors to display low variance, in line with previous work on AE (Cummins & Port, 1998; Tilsen, 2009).

2 Methodology

2.1 Participants, Materials, and Procedures

Participants. We recruited six native Thai speakers with (M) and without musical background (NM). All M participants obtained at least a bachelor's degree in music, while NM had no formal musical training. None of them disclosed any speech or hearing impairment. The presently limited number of participants is due to the demographics of interest, professional musician and musically naïve speakers, and the long experimental duration requiring approximately 2.5–3 h for a full session.

Speech materials. Following previous work on SC, we used ten short phrases with identical prosodic structure "N_1 jùː Prep(osition) N_2" ("N_1 is in/on N_2"). Following (Cummins & Port, 1998), all words in the sentence were monosyllabic and all N_1 and N_2 nouns started with a voiced stop onset to facilitate p-center location. Since *jùː* and *Prep* are function words, they are produced as unstressed.

Procedures. Inside a recording studio, participants sat in front of a computer monitor running a custom GUI used to run the experiment and record audio. Participants were instructed to produce a sentence displayed on the screen and align the first word to a H tone and the last word to a L tone. The H and L tones were generated using a pure tone at 1200 Hz (H) and 600 Hz (L). Tones lasted 50 ms, with 10 ms fade in and out. The H-L interval was kept constant at 700 ms, while the time from the L to a following H was varied so that the H-L interval covers the range 0.3 and 0.75 of the H-to-H PRC, in .05 steps yielding 10 phase values, Figure 3.

In each trial, a random phrase and phase were selected. Participants were instructed to listen to four pairs tones to prepare. Then, participants repeated target sentences ten times aligned with the tones and another ten without the tones while trying to maintain the same phase. We obtained a total of 100 trials (10 unique sentences x 10 phases) per participant and 18- repetition per trial for a total of ~10800 tokens.

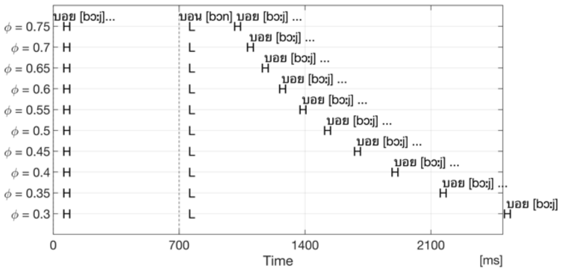

Figure 3: Illustration of rhythmic continuum for stressed syllable alignment in the PRC.

2.2 Data Processing

Following previous work on SC, the main dependent variable we extracted is where the onset of the final stressed word occurs within the phrases in terms of proportion of the phrase. This is also known as observed phase (φ), Figure 4. As a concrete example of a phrase, we calculated the location of the final word p-center (e.g., บอน [bɔːn] in บอยอยู่ในบอน [bɔːj jùː naj bɔːn]) relative to the PRC, which starts and ends with the initial stresseed word of each sentence repetition (e.g., บอย [bɔːj] in บอยอยู่ในบอน… บอยอยู่ในบอน… [bɔːj jùː naj bɔːn…bɔːj jùː naj bɔːn]), Figure 3. Note that the measure is not based on the external rhythmic cue but on participants' productions.

Following the original experiment (Cummins & Port, 1998) and much previous work in the rhythm literature, word onset is not defined as a segmental boundary, but rather as the p-center associated with each word, an event where people perceive prominence and align finger tipping corresponding to maximal change in energy of the signal amplitude envelope.

P-centers were algorithmically located as the midpoint of local rises in the amplitude envelope as follow. To obtain a smooth amplitude envelope that preserves maximally the vocalic energy for each trial, we first down sampled the audio by a factor of 4 to have a frequency of 11025 Hz. We then filtered the signal using a passband first-order Butterworth filter with cutoff frequencies at [700, 1300] Hz. The resulting signal was rectified by taking its absolute value. This procedure was followed by a second round of filtering using a lowpass first-order Butterworth filter with a cutoff frequency of 10 Hz. Finally, we smoothed the amplitude envelope twice using a moving average filter based on 5 samples. To locate the midpoint of rises we started by finding local peaks in the amplitude envelope, rescaled between 0 and 1. We then located the minimum preceding each peak as the closest zero crossing in the gradient of the envelope. Finally, the p-center of each syllable was identified as the midpoint between each minimum and peak, Figure 4.

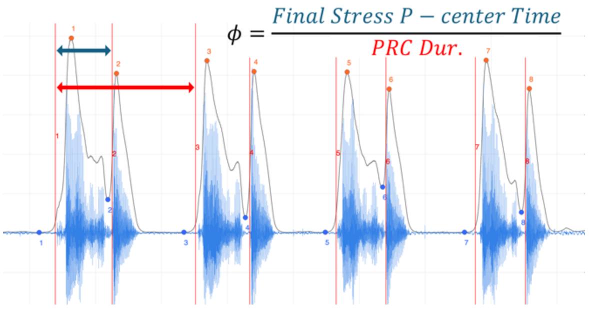

Figure 4: Example of automatic p-center extraction of initial and final syllables for four repetitions and calculation of observed phase (φ). Blue lines mark waveform, gray lines the rectified smoothed amplitude envelope and red line mark detect p-centers between minima (blue dots) and maxima (orange dots).

The phase of each repetition in a trial was calculated as the time of the p-center of the stressed final word divided by the total duration of the PRC spanning the lag between initial p-centers of successive repetitions, Figure 4. Based on a visual display of the amplitude envelope, the locations of peaks, minima, the p-centers were inspected and manually corrected when necessary. Note that, unlike previous work (Cummins & Port, 1998), we did not take the median of all repetitions in a trial, as this could reduce variability and statistical power. Instead, we used all repetitions in all trials. Following previous work (Cummins & Port, 1998), we collapsed repetitions with and without metronome tones, as we observed no significant effects after preliminary testing.

2.3 Data Analysis

Following (Cummins & Port, 1998), we tested the existence of rhythmic constraints in Thai using Gaussian Mixture Models (GMMs) to model the phase distribution both pooling data across subjects and within each subject separately, we tested up to six mixtures and chose their optimal number using the Bayesian Information Criterion (BIC).

To test possible differences between M and NM participants, we obtained bootstrapped 95% confidence intervals for the median of observed phases as a function of target phases. We also fit nested linear mixed-effect regression models to test whether target phase, musical experience, and their interaction are significant predictors of observed phase. All models had by-subjects random intercepts and slopes for musical experience and target phase. Target phase was z-score normalized.

To test whether observed phase is lower at some target phases, separately by subject, we obtained 95% confidence intervals for median values of the interquartile ranges (IQR) and we also fit smoothing splines to the IQR values.

3 Results

3.1 Rhythmic warping: data pooled across all participants

By fitting GMM to observed p-center phase across all participants, we found that Thai speakers warp the rhythmic continuum they are exposed to, Figure 2, into a small number of categories of possible stress locations, Figure 5. These are best modelled with five Gaussian mixtures (μ = .35, .42, .52, .62, .64) capturing three evident modes that gravitate around 1/3, 1/2, and 2/3 of the PRC, Figure 5. In this respect, Thai speaker closely mirror the behavior reported for other languages, like AE.

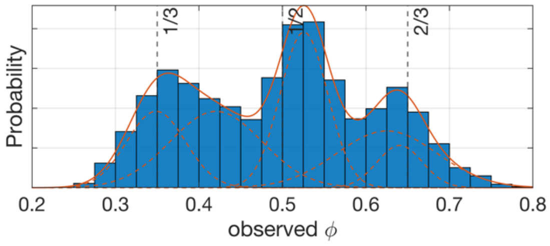

Figure 5: GMM fit to observed relative phases across all participants.

3.2 Rhythmic warping: the role of musical training

Unlike what has been reported for AE, we observed marked differences among participants with and without musical training. A clear difference between M and NM is that they differ in the number of modes displayed in their observed phase. M participants display 3 modes roughly at .33, .5, and .66. NM participants display only 2 modes: .34-.38 and .46-.56, Figure 6.

The rationale for this difference is that M participants can better imitate the phases where the W_1–W_3 group occupies 2/3 (.66) of the phrase repetition cycle. This is illustrated by the 95% CI of the median distance from target phases that is almost invariably < .1 for M and > .1 for NM, when φ > .66, Figure 7.

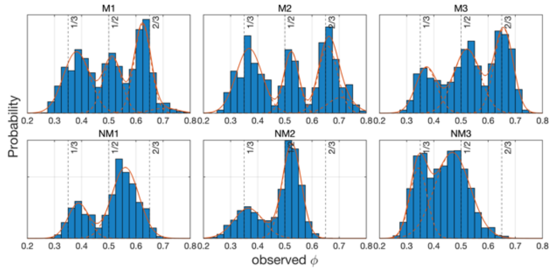

Figure 6: GMM of observed relative phases by participant (top: M, bottom: NM).

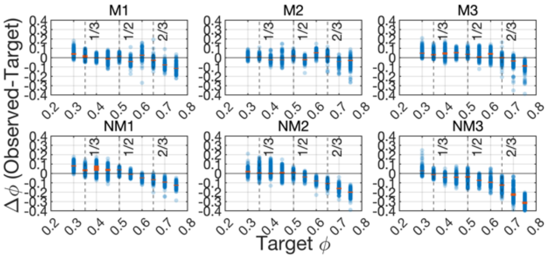

Figure 7: Distribution of distances from target phases with overlaid bootstrapped 95% CI for the median

The different behavior of M and NM is confirmed by fitting LME regressions to their observed phases. The model that best fits the data includes an interaction between target phase and musical experience ($\chi_{(1)}$=13.78, p < .001). The model fit, Figure 8, shows that observed phase increases with target phase, indicating that participants correctly perform the task. Intercept for .5 phase is 0.53 (95% CI [0.52–0.55], p < .0001). Observed phase increases by .11 (95% CI [0.10 –0.13], p < .0001) per .15 increase in target phase, indicating a close match. Lack of musical experience is associated with a lower intercept (−0.04 95% CI [−0.08 −0.001], p = .04) indicating that there the match of target phases is less accurate even at .5 target phase for NM participants. NM participants struggle more to match target phases as target phase increases, as reflected in a negative interaction between lack of musical experience and target phase (−0.05 95% CI [−0.072 −0.032], p < .0001).

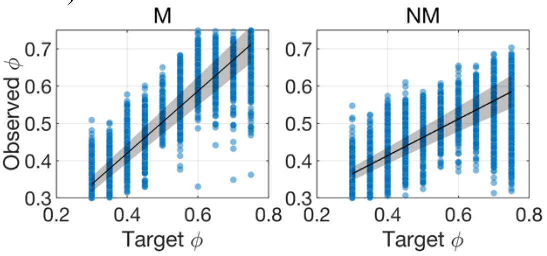

Figure 8: Model fit of observed phase as a function of target phase for M (left) and NM (right).

3.3 Rhythmic warping: dynamical systems' signatures

Finally, we studied variability based on bootstrapped 95% CI of the median IQR obtained using all repetition in a trial. From this analysis, the dynamical signature of lower variance in and around integer ratios (1/3, 1/2, and, to a lesser extent, 2/3) of the PRC also emerges, Figure 9 Top. This fact is reflected in the "dips" in the smoothing spline fits, Figure 9 Bottom. Low variability around ~.33 is evident for M1, M2, M3, NM1, and NM3. Low variability around ~.5 is exhibited by all participants. Finally, low variability around ~.66 is less clear, but seems present for M1, M2, and NM1.

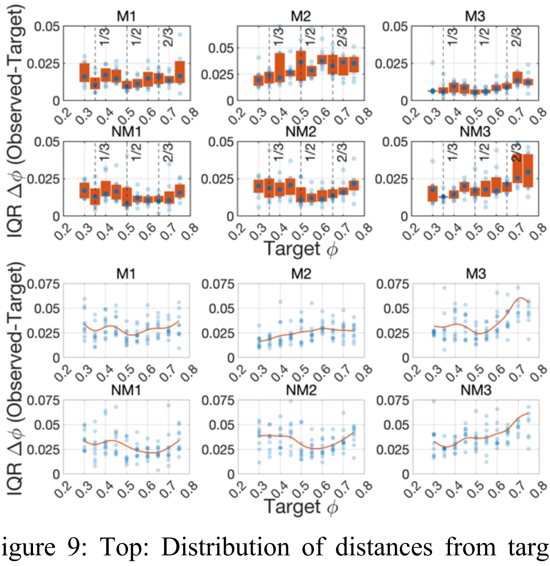

Figure 9: Top: Distribution of distances from target phases IQRs with overlaid bootstrapped 95% CI for the median. Bottom: smoothing splines fit to the same data.

4 Discussion and Conclusion

We now return to our research questions. The first question we investigated is whether Thai speakers exhibit rhythmic constraints in their production of stressed syllables. The GMM modeling strongly suggests that they do, as Thai speakers are highly constrained in their placement of final stressed syllables within a phrase. When exposed to a continuum of possible alignments for stressed syllables within a phrase, Thai speakers, with or without musical experience, cannot fully reproduce that continuum. To the contrary, they warp it into a small number of alignment categories that tend to divide the PRC into simple integer ratios such as 1/3, 1/2, 2/3. This finding replicates previous work on languages like AE, cited in the introduction. Our findings allow us to further substantiate the claim that the impression of rhythmicity in speech may come from higher level constraints on production and perception. These constraints dictate a small number of possibilities for the placement of prominent events, like stressed syllables. This limited range of possibilities for the placement of prominent events, in turn, can give rise to an impression of rhythmicity. This impression of rhythmicity arises simply from the fact that only a limited number of patterns is available, and, thus, the patterns end up being repeated, leading to an impression of periodicity and rhythmicity.

The second question we investigated is whether there are individual differences in rhythmic warping behaviors based on rhythmic/musical training. Recall that previous work on AE (Cummins & Port, 1998) found no differences between participants that do and do not have such training. This finding was taken as evidence for the cognitive and universal nature of the constraints at play in production/perception.

However, our combined evince from GMM fitted by–participant and linear mixed effect regression shows that a clear difference exists between participants with and without musical training in Thai, unlike in AE. Thai participants with a musical background display a low (1/3), mid (1/2), and high attractor (2/3), while participants without a musical background display only the first two attractors. This is an important aspect that may have been overlooked in previous work, as it shows that rhythmic constraints are not purely cognitive in nature, but they also stem from linguistic experiences – as shown by the difference between AE and Thai – and by individual experience – as shown by the difference between participants with and without musical training.

We now briefly discuss how the differences in performance between Thai and AE speakers may be related to the phonological properties of the two languages. Unlike English varieties, Thai is a language where rhythm is iambic and prominent elements are group final (Bee, 1975; Bennett, 1994). Thus, for example, the grouping of a phrase would be [ˈbɔːj] [jùː naj ˈbɔːn] vs. [ˈbeg for a] [ˈdime] in English (Cummins & Port, 1998). Note that this grouping with final prominence is expected in Thai not only on the grounds of phonological analyses, but also of Thai traditional music grouping.

Keeping in mind this background, we can attempt to relate the attractors we observed to low dimensional phonological analyses. For attractors at l/3 and 1/2 of the PRC, participants produced two stress groups [ˈbɔːj] [jùː naj ˈbɔːn] with final prominence and considerably shorter W_2 and W_3 compared to W_1 and W_4, that is [–] [◡◡–]. The only difference between attractors at l/3 and 1/2 is the presence of a silent beat in the 1/3 case [–] [◡◡–] [ ] (Cummins & Port, 1998).

However, a different strategy must be adopted when an attractor is displayed at 2/3 of the PRC. Specifically, three stress groups are needed and, in AE, this requires introducing a stress on words that are normally unstressed like *for*, *i.e.*, [ˈbeg] [ˈfor a] [ˈdime] (Cummins & Port, 1998). In Thai, speakers also need to produce three stress group, as, e.g., [ˈbɔːj] [jùː ˈnaj] [ˈbɔːn] [–] [◡–] [–] on normally unstressed words. This is exactly what M participants do and NM participants fail to do.

We believe that a potential reason for this failure is that Thai speakers tend to normally create trisyllabic stress groups constituted by a single anapestic foot in faster speech ([◡◡–]). A grouping where most of the duration is concentrated on the final prominent syllable. However, this strategy is incompatible with introducing stress on either of the preceding two words, as is necessary for the final stressed syllable to appear at 2/3 of the PRC.

Our hypothesis is based on previous phonological and experimental work. For Thai speakers, a tendency towards polysyllabic stress groups ending in a longer prominent syllable has been reported in the phonological literature (Rudaravanija, 1965) and also experimentally confirmed (Nitisaroj, 2004; Potisuk et al., 1996).

Moreover, syllable durations that are in line with anapestic rhythm ([◡◡–]) have been reported as the routine realization of trisyllabic compounds and phrases (Gandour et al., 1992).

In sum, NM participants display only patterns that seem in line with what has been observed in non-rhythmic Thai speech. M participants, on the other hand, can produce other less common patterns. In our opinion, the difficulties manifested by speakers without musical training in producing a rhythm compatible with an attractor at 2/3 of the PRC could be an additional manifestation of the strongly iambic rhythm of Thai. A property that sets this language apart from other languages like AE and that makes Thai run contrary to a more "isosyllabic" or syllable-timed rhythm often observed at high speech rates (Arvaniti, 2012).

The final question we have investigated is whether the rhythmic constraints we reported may betray the signatures of an underlying dynamical system of coupled oscillator that could be used as a way to conceptualize the observed rhythmic patterns, as hypothesized in recent work (Nam et al., 2008; O'Dell & Nieminen, 1999; Tilsen, 2009).

To test this question, we supplemented GMM models with analyses of variability. Our analyses of variability at different phase values using both bootstrapping of IQRs and spline smoothing confirms previous findings of lower variability around the centers of the categories in which the rhythmic continuum of stress placement is warped by participants.

The combined findings of a warping of the rhythmic continuum into a small number of more stable categories and the lower variability of said categories are compatible with previous suggestions, (e.g., Port, 2003), that, in SC, rhythm can be generated by two coupled oscillators of different frequencies for the stress groups (or metrical foot) and the phrase. These oscillators evolve according to a potential function representing the phase of the slower PRC when a new stress group is initiated (Port, 2003).

For reasons of space we refrain from presenting a full discussion of a coupled oscillator model of speech cycling in Thai. Yet we wish to point out that we have developed a computational implementation of the dynamical system proposed in previous work and, by further parametrizing a previous proposal (Port, 2003), we found that it qualitatively mirrors well our data pooled across participants, Figure 10.

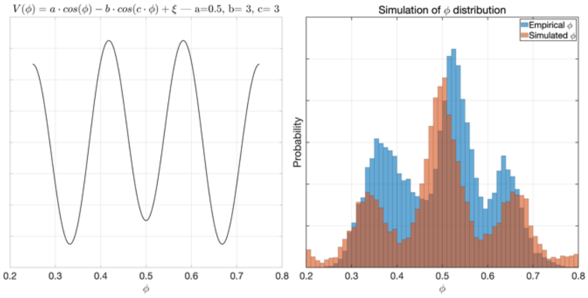

Figure 10: Left: potential function of phase dynamics. Right: stochastic dynamical system model simulation vs empirical data.

As shown in Figure 10, this dynamical system correctly captures macroscopic properties of the constraints observed on the placements of stressed syllables in Thai. That is to say, it captures a strong tendency for stressed syllables to be produced around 1/3, 1/2, and 2/3 of the phrase.

In addition, tuning the parameters a, b, and c of the model in the equation in Figure 10 allows us to generate cross-linguistic and cross-individual variation. Minimal changes to these parameters can, for example, generate behaviors where only two attractors are present, as we have observed for some of participants with no rhythmic training, e.g., NM3 in Figure 6, as illustrated in Figure 11.

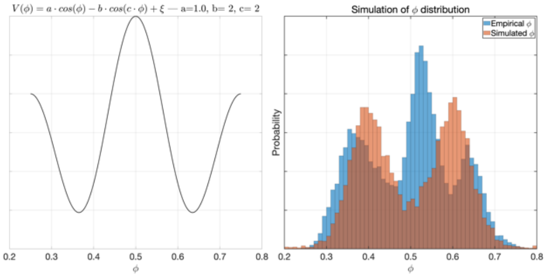

Figure 11: Left: potential function of phase dynamics. Right: model simulation vs empirical data.

We can thus think of said parameter modulations as a capturing how rhythm is the result of a universal laws (the dynamical system potential) and parameters that may be modulated by a variety of factors such as linguistic background (for instance being a speaker of English or Thai) or individual experiences (for instance having musical training, auditory acuity, previous experience with rhythmic tasks, *etc*).

To conclude, this study shows that Thai speakers exhibit rhythmic constraints in speech, aligning stressed syllables at simple ratios integer like 1/3, 1/2, and 2/3 of a phrase. This finding indicates that universal cognitive processes can give origin to the impression of rhythm in speech because the number of available patterns is limited, thus repetition becomes the norm. Importantly, musical training can loosen these constraints. Musically trained speakers show more distinct rhythmic patterns than speakers without such training. This second finding suggests that, while some rhythmic constraints are universal, others are shaped by individual experience. Finally, our findings also reveal how Thai's iambic, prominence-final rhythm interacts with these constraints, with non-musically trained speakers reflecting natural speech patterns more closely. We have proposed that this dual nature of constraint on speech rhythm can elegantly be capture by a dynamical system. In this system, the potential function reflects universal tendencies that are further modulated by parameter modulations capturing individual experience.

Thus, our results support a view of speech rhythm as the manifestation of a complex interplay between cognitive mechanisms and individual experience that shape speech behavior in language- and individual-specific ways.

Acknowledgements

The work was supported by a Silpakorn University grant (66.0329-040-3605). Both authors contributed equally to this research project.

References

Abercrombie, D. (1990). *Elements of General Phonetics*. Edinburgh University Press. https://doi.org/doi:10.1515/9781474463775

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. *Journal of Phonetics*, *40*(3), 351–373.

Bee, P. (1975). Restricted phonology in certain Thai linker-syllables. *Studies in Tai Linguistics in Honor of William J. Gedney. Bangkok: Central Institute of English Language*.

Bennett, J. F. (1994). Iambicity in Thai. *Studies in the Linguistic Sciences*, *24*(1), 39–57.

Bertinetto, P. M. (1985). A proposito di alcuni recenti contributi alla prosodia dell'italiano. *Annali Della Scuola Normale Superiore Di Pisa. Classe Di Lettere e Filosofia*, *15*(2), 581–643.

Bertinetto, P. M. (1989). Reflections on the dichotomy 'stress' vs.'syllable-timing.' *Revue de Phonétique Appliquée*, *91*(93), 99–130.

Bertinetto, P. M., & Bertini, C. (2010). Towards a unified predictive model of Natural Language Rhythm. In *Prosodic Universals. Comparative studies in rhythmic modeling and rhythm typology.* (pp. 43–78). Aracne.

Chung, Y., & Arvaniti, A. (2013). Speech rhythm in Korean: Experiments in speech cycling. *Proceedings of Meetings on Acoustics*, *19*(1).

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

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*(1), 51–62.

Dauer, R. M. (1987). Phonetic and phonological components of language rhythm. *Proceedings of the 11th International Congress of Phonetic Sciences*, *5*, 447–450.

Dellwo, V. (2006). *Rhythm and speech rate: A variation coefficient forΔ C. InKarnowski, Pawel & Szigeti, Imre (eds.), Language and language-processing, 231–241*. Frankfurt: Peter Lang.

Fletcher, J. (2010). The Prosody of Speech: Timing and Rhythm. In *The Handbook of Phonetic Sciences* (pp. 521–602). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444317251.ch15

Franich, K. (2021). Metrical prominence asymmetries in Medumba, a Grassfields Bantu language. *Language*, *97*(2), 365–402.

Gandour, J., Dechongkit, S., Ponglorpisit, S., & Kim, S. Y. (1992). Intraword Timing Relations in Thai. *Pasaa*, *22*.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology*, *7*(515–546).

Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. University of Chicago Press.

Kohler, K. J. (2009). Rhythm in Speech and Language. *Phonetica*, *66*(1–2), 29–45. https://doi.org/doi:10.1159/000208929

Luangthongkum, T. (1978). *Rhythm in Standard Thai*. The University of Edinburgh.

Mairano, P., & Romano, A. (2011). Rhythm metrics for 21 languages. *Proc. of the 17th International Congress of Phonetic Sciences*, 1318–1321.

Nam, H., Saltzman, E., Krivokapić, J., & Goldstein, L. (2008). Modeling the durational difference of stressed vs. Unstressed syllables. *Proceedings of the 8th Phonetic Conference of China*.

Nitisaroj, R. (2004). Perception of stress in Thai. *The Journal of the Acoustical Society of America*, *116*(4_Supplement), 2645–2645.

O'Dell, M., & Nieminen, T. (1999). Coupled oscillator model of speech rhythm. *Proceedings of the XIVth International Congress of Phonetic Sciences*, *2*, 1075–1078.

Pantupong, W. (1973). Pitch, Stress and Rhythm in Thai. *Pasaa*, *3*(2), 41–62.

Pike, K. L. (1945). *The intonation of American English* (Vol. 1). University of Michigan Press.

Port, R. F. (2003). Meter and speech. *Journal of Phonetics*, *31*(3–4), 599–611.

Potisuk, S., Gandour, J., & Harper, M. P. (1996). Acoustic Correlates of Stress in Thai. *Phonetica*, *53*(4), 200–220. https://doi.org/doi:10.1159/000262201

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292.

Rudaravanija, P. (1965). *An analysis of the elements in Thai that correspond to the basic intonation patterns of English*. Columbia University.

Suntornsawet, J. (2022). A Systemic Review of Thai-Accented English Phonology. *PASAA: Journal of Language Teaching and Learning in Thailand*, *63*, 348–370.

Tajima, K., & Port, R. F. (2003). Speech rhythm in English and Japanese. *Phonetic Interpretation: Papers in Laboratory Phonology VI*, 317–334.

Tilsen, S. (2009). Multitimescale dynamical interactions between speech rhythm and gesture. *Cognitive Science*, *33*(5), 839–879.

Turk, A., & Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Laboratory Phonology*, *4*(1), 93–118.

Zawaydeh, B. A., Tajima, K., & Kitahara, M. (2002). Discovering Arabic rhythm through a speech cycling task. *Perspectives on Arabic Linguistics XIII-XIV: Papers from the Thirteenth and Fourteenth Annual Symposia on Arabic Linguistics*, *230*, 39.