QUANTIFYING VOWEL ONSET PERIODICITY IN JAPANESE

Michael Connolly Brady *Robert F. Port*

Departments of Linguistics and Cognitive Science, Indiana University
330 Memorial Hall, Bloomington, Indiana 47405
mcbrady@indiana.edu port@indiana.edu

ABSTRACT

Many researchers will agree that the brain must employ some expectancy mechanism for speech as speech unfolds through time. We, among others, posit that an adaptive oscillator may be made to synchronize with these speech expectancies. The oscillator would define something of a cognitive pulse and should align itself with key features of the acoustic signal. A recent study we conducted on compensatory mora relationships between neighbor voicing inter-onset-intervals in Japanese strongly indicates voice onsets as important targets for a referent pulse or planning mechanism. We review that study and draw on its results to highlight some issues related to modeling a referent or cognitive pulse. Based on a circular statistics foundation, we note how some vowel onsets should be treated as strong coupling targets for an adaptive oscillator while other vowel onsets should be treated more as distractions. From there we discuss some problems and issues associated with analyzing speech for forms of periodic regularity.

Keywords: mora timing, Japanese speech rhythm, temporal compensation, beat induction.

1. INTRODUCTION

Japanese has long been claimed to have temporal regularity in its pronunciation based on *mora* units. A mora is archetypically a consonant-vowel syllable but it can also be a long vowel or long consonant. The graphemes in the kana writing system correspond to mora and each represent either a CV syllable or a lengthening feature for the preceding vowel or preceding consonant. Japanese pedagogy holds that each mora takes the same amount of time and this amounts to assigning the same amount of time to each symbol in kana notation. For instance, a word like *To-mi-ko* should take the same amount of time as *To-o-ko* or *To-k-ko* or *Ta-n-ka*, where dashes separate spaces between kana graphemes.

After decades of research, controversy remains as to whether there is any inherently temporal nature to the mora. Some researchers [4, 9, 10, 15] argue that moras reflect a global coordination of speech timing while others [3, 17, 18] argue that Japanese syllables are simply pronounced sequentially with no explicit global coordination. Here, the approximate timing regularity that is observed is claimed to fall out from the natural timing of each individual unit and temporal variance should accumulate. This is described as the *accumulative variance* hypothesis by Warner and Arai [17, 18]. We take the former perspective and argue that Japanese is coordinated in larger mora-sized chunks and find evidence for this in the form of temporal compensation between neighbor speech segments. Durations of segments seem to stretch and compress so as to align with some production control mechanism that supports global regularities. We refer to this as the *temporal compensation* hypothesis [15]. A wealth of literature continues to mount in the debate between these competing perspectives.

In our research on this issue, we have come to redefine the relevant time interval. This interval is the one whose duration is most systematically realized [1, 7, 16] as the interval between vowel onsets. We call our interval the 'vowel-onset mora' or 'Vmora'. *This new theoretical unit contrasts with the traditional mora unit in that it diverges from kana notation.* For instance, the traditional units *ta ka gi to*.. might be more aptly notated as *(t) ak ag it*.. in our system. As with the traditional notion of the mora, Vmoras should be separated from each other by 1, 2, 3 etc. mora-timed cycles. Our rationale for this revision comes from data that show vowel onsets to have a tendency to be spaced at near integer ratios [9, 10, 15]. In this paper, we present some of our recent evidence that preserves the regularity of vowel onsets in Japanese. We go on to shed further light on what is happening with Japanese speech timing as we discuss some methods and issues related to analyzing speech for periodicity and adaptive periodicity.

2. PROCEDURE

We made recordings of four female monolingual Japanese speakers of the Kanto (Tokyo) dialect in four speech production tasks to look for effects of speaking style. The tasks involved reading text and spontaneously talking about people in situations from a set of illustrations. In the first two recording conditions speakers were asked to read from flashcards where a flashcard had a carrier sentence printed on it that contained one of six target words. These target words were the names of characters, all with the same family name, *Takagi,* and varying personal names, *Toko, Tooko, Tokko, Tomiko, Toshiko* and *Tonko*. The carrier phrases and names were written on the flashcards using hiragana, katakana and kanji, the standard Japanese orthography. An example carrier sentence is presented in Table 1. Carrier sentences were designed to be either formal or informal and the speakers were urged to read the cards in the appropriate style. There were three carrier sentences per speaking style and our four speakers read through three sets of flashcards per style for a total of 432 recordings (6 target words x 3 carrier sentences per style x 2 speaking styles x 3 recordings per condition x 4 speakers = 432).

In the third or spontaneous recording condition, speakers were interviewed about people and situations using a series of comic-book style illustrations. The names of characters were printed under the illustrations, using 3 of the 6 names above (*Tomiko, Tooko, Tokko)*. A Japanese experimenter and the speakers discussed the characters and situations in the illustrations while the experimenter worked to get the speakers to spontaneously say the full character names. Once the speaker was recorded using each of these names, the experimenter moved on to the next picture. If necessary, the experimenter would probe with questions like "which person do you think is the most stylish?" Or "who is the older sister?" The procedure continued until the experimenter had elicited nine utterances of each of the three target characters' full names from a speaker (9 x 3 x 4 = 108 recordings).

Our design for the first three conditions compares the timing of two- and three-mora units of Japanese that are known to pose difficulties for mora regularity [15]. Explicitly, a word like *To-mi-ko* has three syllables and three moras while *To-k-ko* and *To-o-ko* also have three moras but only two syllables. These two-syllable words tend to not fill

Table 1: Sample carrier sentence and target words.

あちらのテーブルの一番左の方は人事部の高木_____さんです。
(Romaji: Achira no teeburu no ichiban hidari no kata wa jinjibu no Takagi _____ san desu)
(English: The left-most person at the table over there is Ms. _____ Tagaki in the personnel department)
* Blanks filled with one of: Toko, Tokko, Tooko, Tonko, Toshiko, Tomiko

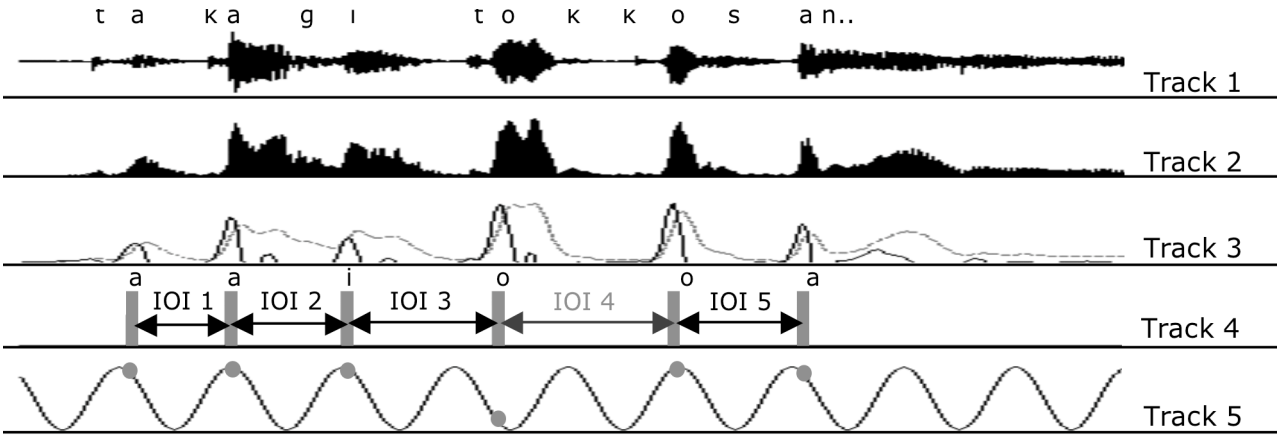
their theorized three-mora time allotments because the mora contributed by the geminate stop (-k-) or long vowel (-o-) are typically not as long as a regular CV mora (-mi-). The hypothesis then is that geminates and long vowels should illicit compensatory lengthening by neighboring Vmoras if a full three-mora word is to be produced.

In a final condition, speakers were asked to read a set of three phrases as if their recordings would be used to train beginning students of Japanese. The phrases were chosen from a Japanese phrase book for learners. Our speakers repeated each phrase a number of times until the experimenter was confident that a well-rehearsed recording of each phrase was obtained in a clear and formal style.

2.1. Analysis

We extracted segments containing the full names of the characters, including the honorific suffix *san,* from each of the formal, informal, and spontaneous recordings as tokens for analysis. Consequently, all tokens had six vowel onsets where the interval between the fourth and fifth vowel onsets served as the interval we varied via the characters' personal names. We looked for compensation between inter-onset-intervals (IOIs) depending on the Vmora pattern of each token. As for the effects of speaking style, we found differences in overall rate of speech (the informal style was spoken a little faster than the formal or

Figure 1: Automatic analysis for a typical token of "Takagi Tokko san".



spontaneous styles), but analysis found no differences in durational ratios between the formal, informal or spontaneous speaking conditions. Thus we pool over those conditions.

Fig. 1 illustrates a typical token and its processing. Track 1 of the figure presents the original recording with its phonetic labels. Track 2 presents its amplitude envelope, a low-pass filtered and rectified version of the signal. Track 3 displays the positive rate of change of a smoothed version of the amplitude envelope, found at 5 ms intervals. A peak-finding algorithm determined the locations of maximal attack whose peaks define vowel onsets, as depicted with vertical bars in Track 4. The inter-onset-intervals (IOIs) are the spaces between vowel onsets as labeled in Track 4.

Each token was visually inspected to check the extracted onsets for accuracy and all suspect tokens were discarded. For example, the *-g-* in *Takagi* was sometimes not clearly articulated so the *-i-* onset was sometimes either missed or was detected in a questionable location. Of the 540 total tokens, 391 (72%) of them were used for the analysis.

2.2. Results: Temporal Compensation

To normalize tokens for speech rate so that tokens could be compared with each other, we computed a Reference Period (RP) for each token and divided IOIs of the token by the RP. The RP for a token is taken as the ratio of the duration of the token (measured from first vowel onset to last vowel onset) to the supposed total number of moras in the token. This supposed number is usually six but is five in the case of *Toko*. The Reference Period of a token can be thought of as an estimate of the cycle period of the hypothetical 'mora oscillator' for the token. Track 5 of Fig. 1 depicts a sinusoid with a period that was calculated as the token's RP.

In Fig. 2, the IOIs from Track 4 of Fig. 1 are rotated vertically and scaled by their Reference Period. Since for each speaker and style individually the durational patterns look like Fig. 2 in all essential respects, the data shown are pooled across the four speakers and three speaking styles for each target word. Here we see that regardless of target word, IOI-1, IOI-2, and IOI-5 of the carrier phrase are very close to each other in duration and to the standard RP. IOI-3 and IOI-4 are more difficult. In each case (except *Toko*) the fourth 'contains' two Vmoras while the other IOIs have one Vmora. IOI-4 values fall around 1.65 (for *Tokko* and *Tooko)*, for instance and around 1.85 (for *Toshiko* and *Tomiko)*, well short of the 2.0 RP that might be expected. However, notice that the IOI-3 means vary inversely with their corresponding IOI-4 means and the sum of the two intervals yields a number close to three. This can be seen most clearly in Fig. 3 where each IOI-3 is plotted against IOI-4 for all the two-Vmora IOI-4s in the experiment. The line IOI-3 + IOI-4 = 3.0 RP (solid) fits nicely through the data points. The best-fitting regression line (dashed) through the data lies very close to this line. This clearly demonstrates strong compensatory timing between the two intervals. In fact, one might say that only 2/3 of the additional time for the lengthened Vmoras (in *Tooko* and *Tokko*) lie within the Target IOI-4, about 1/5 is achieved by lengthening the Anticipatory IOI-3.

The question is raised, what kind of mechanism could (a) treat RP intervals as temporal attractors and (b) compensate for a short IOI by anticipatory lengthening of the preceding IOI.

Figure 2: Mean normalized IOI durations

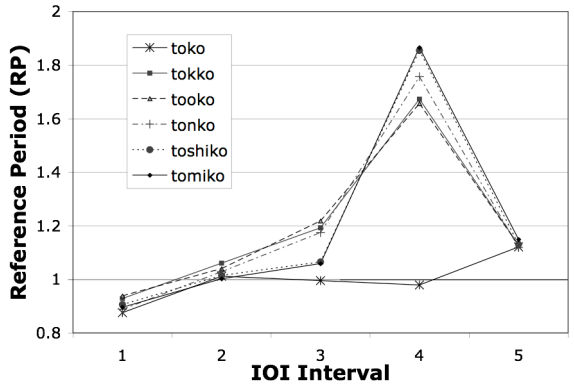
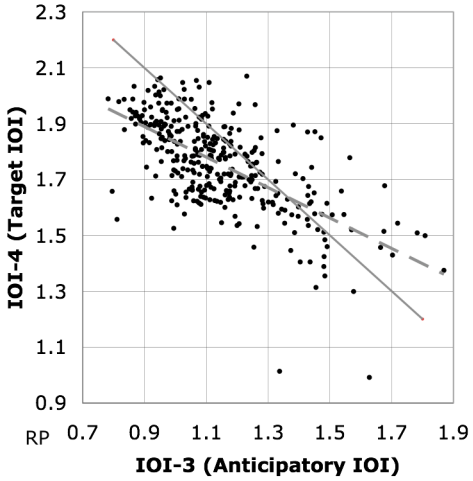


Figure 3: IOI-3 versus IOI-4



3. QUANTIFYING PERIODICITY

With temporal compensation in mind, a method for examining relatively longer patterns of onsets is called for. We need to evaluate whether the IOIs of entire phrases stretch and compress in relation to each other so as to align with a global production or control mechanism. An obvious first kind of temporal organization to look for is periodicity. This section introduces ways for analyzing IOI sequences for such structure.

3.1. Phase Clustering

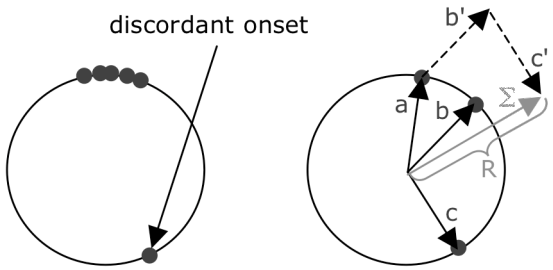
A sequence of onsets may be treated as cyclical data with respect to a simple sinusoid. From this we may apply techniques from circular statistics to analyze how the data clusters with respect to some periodicity. For instance, at the time of an onset, the phase of a sinusoid such as depicted in Track 5 of Fig. 1 may be sampled to provide a data point. Plotting data points on a circle then translates Track 5 of Fig. 1 into Fig. 4a. Here we note that the onset between the 3rd and 4th IOI is discordant in circular statistics terminology. A discordant observation finds itself significantly far away from the mean of the data mass and is not "in harmony" with the other data points in relation to the sinusoid.

A second important term from circular statistics is that of the quantity $\overline{R}$, a measure comparable to the standard deviation of a normal distribution. A *resultant length* R is defined as the sum of the vectors corresponding to data points. The *mean resultant length,* $\overline{R}$, takes a value between zero and one as R is divided by its sample size. Fig. 4b helps to visualize this as the vectors a, b, and c are summed to define a point. Distance to this point from the origin of the unit circle is R. From this we see that if the data points cluster together, their corresponding vectors will mostly point in the same direction and $\overline{R}$ will be relatively large. If the data points are distributed around the circle, $\overline{R}$ will be small. There is an excellent literature on analyzing cyclic data and discordant observations with circular statistics [6, 8].

We now put $\overline{R}$ to use. The accumulative variance hypothesis asserts that a sequence of IOIs will appear to have timing regularity regardless of serial order. If this is true, a sequence of IOIs from an utterance in their original order should appear to have roughly the same timing regularity as a typical random-ordered sequence of the same set of IOIs. We evaluated this prediction.

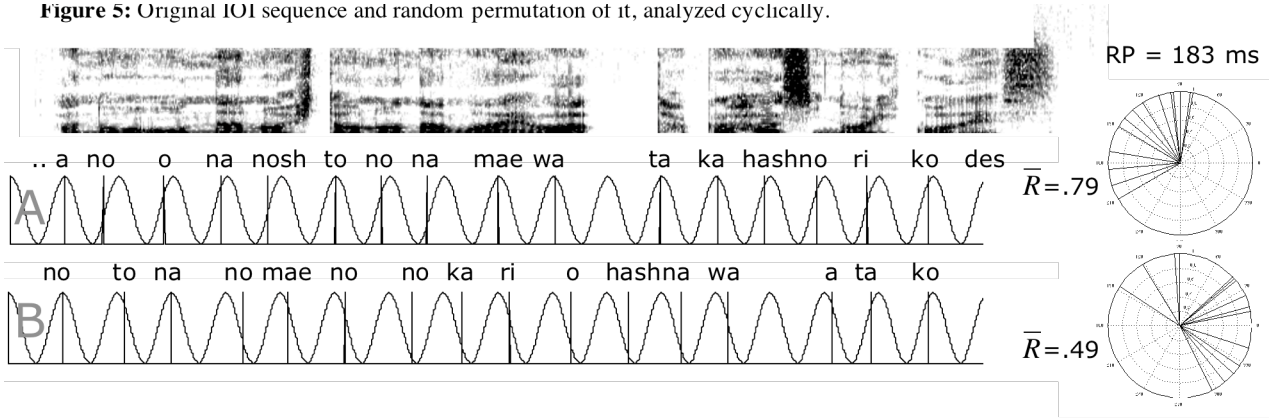
Sequence A of Fig. 5 illustrates an utterance we arbitrarily selected from the phrase book recording condition. We hand-coded its vowel onsets and estimated a Reference Period as the sum of the IOIs divided by 16, the number of IOIs in the segment (counting /wa/ as two intervals). We generated a sinusoid with the period of this RP to

Figures 4: a) Clustering (left), b) calculating R (right).



$$R^2 = \left(\sum_i \sin(2\pi\theta_i)\right)^2 + \left(\sum_i \cos(2\pi\theta_i)\right)^2, \quad \overline{R} = \frac{R}{n}$$

Figure 5: Original IOI sequence and random permutation of it, analyzed cyclically.



map voice onsets into circular data points. We then calculated $\overline{R}$ from those points and found they clustered relatively well ($\overline{R}$=.79). We then took the same set of IOIs, and we randomly shuffled their order. We used the same RP (the sum of the IOIs hasn't changed nor has the number of IOIs) to again map onsets to circular data points. We performed this shuffle-map-compute procedure 50 times to determine a '*mean random* $\overline{R}'$ of .48 for the sequence. A representative shuffle with $\overline{R}$=.49 is presented as Sequence B of Fig 5. Here we find Sequence A is more periodically structured than Sequence B. We mean Fig. 5 to be interpreted as merely an illustration in favor of global temporal compensation. This general result is consistent over the limited number of speech utterances we analyzed, but a thorough statistical analysis over a corpus of speech samples is called for.

3.2. Adaptive Oscillators

Mapping an onset sequence to an RP sinusoid and taking its $\overline{R}$ is much like performing a Fourier transform on the sequence and inspecting the part of the resultant frequency spectrum corresponding to the RP. However, not only does this require putting the sequence into a buffer for analysis, it also is not robust to tempo change. What is needed is a way of adjusting the RP as the sequence unfolds through time. Numerous researchers have developed useful methods for tracking periodic and quasi-periodic sequences by using adaptive oscillators, e.g. [2, 5, 12, 13]. To make our discussion as precise as possible, we introduce a few equations to describe a generic adaptive oscillator. The oscillator's phase and period are updated at the time of an event. Phase, θ, is incremented by the portion of the oscillator's period, *p,* that has elapsed since the last update with:

$$\theta_t = \theta_{t-1} + \frac{IOI}{p} \quad (1)$$

and in order to synchronize with an onset sequence, the period of the oscillator must be modified at the time of each onset with an equation such as[i]:

$$\Delta p = -p + \left(p \cdot \eta \cdot E(\theta)\right) + G \quad (2)$$

where η is a constant that reflects coupling strength. Finally, E is the synchrony error function that specifies how much and in what direction the oscillator should adapt based on its phase:

$$E(\theta) = \sin(2\pi\theta) \quad (3)$$

By operating under the assumption that an onset pattern is isochronous, fluctuations from isochrony can be considered as temporal noise to be filtered out. This assumption amounts to the idea that onsets will be normally distributed (called the von Mises distribtion) about a mean phase on the circle. Here we have a dilemma. The notion of discordant onsets or of a multimodal circular distribution is in direct conflict with the premise that an oscillator is tracking a noisily isochronous sequence. The ideal solution in terms of these equations would be to somehow set η to zero at the time of a discordant event so that the oscillator does not adapt. The problem is that using the oscillator itself to forecast what to couple with and what to ignore eventually leads to catastrophic decisions. Some researchers such as [2, 12] get around this by cleverly incorporating a

second oscillator with a longer period where the two seek to synchronize with each other as well as to the input sequence. Though their approaches exhibit some success, we constrain ourselves to exploring how a lone adaptive oscillator might be implemented to ignore discordant events. The general approach we pursue involves a bank of resonators that are tuned to a range of frequencies [5]. At any given instant, the integrated output of the bank specifies an expectation value. This value is passed through a threshold function to determine η at any given time step.

4. DISCUSSION

This paper has highlighted some issues related to compensatory timing and the search for an adaptive reference periodicity in Japanese speech. In considering the orations of preachers and rappers, most will agree that speech can readily be made to have a rhythmic pulse. Yet, regardless of language, everyday conversational speech rarely seems to exhibit such an obvious pulse. In light of evidence we find for temporal compensation in Japanese, we propose that conversational speech may be riddled with discordant speech cues and that this may be the primary reason most attempts to automatically detect reference periodicities in conversational speech have failed.

The two points we emphasize in summary of this paper are that 1) the durations of speech segments as related to the mora in Japanese depend on the durations of other speech segments, and 2) this compensation will foil any adaptive oscillator that assumes onsets to be noisily isochronous.

We lastly point out that an adaptive oscillator or cognitive or reference pulse is merely an abstraction that allows us to talk about expectations in speech. A cognitive oscillator may or may not be realized as some cohesive neural mechanism. To us it is a device for orchestrating interactions between memories and expectations in the mind where the concept of serial structure becomes fuzzy. Tongue twisters and spoonerisms and the phenomenon of phonemic restoration all lend credence to the notion that an utterance somehow exists *simultaneously* in active memory. As the words of an entire utterance must concurrently be active to contextualize each other, the utterance must also be converted into coordinated sequences of articulations. The idea of an adaptive oscillator provides the referential structure needed to convert abstract and interdependent mental representations into the serial stream of speech.

5. REFERENCES

[1] Allen, G. 1975. Speech rhythm: Its relation to performance universals and articulatory timing. *Journal of Phonetics* 3, 75–86

[2] Barbosa, P. 2007. How prosodic variability can be handled by a dynamical speech rhythm model. *Proc. 16th ICPhS* Saarbrücken, (this issue)

[3] Beckman, M. E. 1982. Segment duration and the 'mora' in Japanese. *Phonetica* 39, 113-35

[4] Brady, M.C., Port, R. F., Nagao, K. 2006. Effects of speaking style on the regularity of mora timing in Japanese. *J. Acoust. Soc. Am.* 120(5), 3208

[5] Brady, M.C. (in preparation). Adaptive beat tracking for non-isochronous rhythms

[6] Collett, D. 1980. Outliers in circular data. *Appl. Statist* 29, 50-57

[7] Cummins, F., & Port, R. 1998. Rhythmic constraints on speech timing. *Journal of Phonetics* 26, 145–171

[8] Fisher, N. I. 1993. *Statistical Analysis of Circular Data*. New York : Cambridge University Press

[9] Han, M.S. 1994. Acoustic manifestations of mora timing in Japanese. *J. Acoust. Soc. Am*. 96, 73–82

[10] Hirata, Y., Whiton, J. 2005. Effects of speaking rate on the single/geminate stop distinction in Japanese. *J.Acoust. Soc. Am* 118 (3), 1647-1660

[11] Kato, H., Tsuzaki, M., Sagisaka, Y. 1998 Acceptability for temporal modification of single vowel segments in isolated words. . *J. Acoust. Soc. Am*. 104, 540-549

[12] Large, E. W., Jones, M. R. 1999. The dynamics of attending: How we track time varying events. *Psychological Revie*w 106, 119–159

[13] McAuley, D. 1995. Perception of time as phase: Toward an adaptive oscillator model of rhythmic pattern processing. Unpublished doctoral dissertation, Indiana University, Bloomington

[14] Lehiste, I. 1977. Isochrony reconsidered. *Journal of Phonetics* 5, 253-263

[15] Port, R., Dalby, J., O'Dell, M. 1987. Evidence for mora timing in Japanese. *J. Acoust. Soc. Am.* 81, 1574-85

[16] Port, Robert. 2003. Meter and speech. *Journal of Phonetics* 31, 599-611

[17] Warner, N., Arai, T. 2001. Japanese mora-timing: A review. *Phonetica*. 58, 1–25

[18] Warner, N., Arai, T. 2001. The role of the mora in the timing of spontaneous Japanese speech. *J. Acoust. Soc. Am.* 109(3), 1144-1156

[i] G is a term that can involve a variety of adaptive components. For instance one component may create tension on the oscillator as a tendency for it to decay towards a preferred period. Or, G might only involve a phase shift distributed over a single cycle, such as:
$G = p_t \eta_\theta E_\theta(\theta_t) - p_{t-1} \eta_\theta E_\theta(\theta_{t-1})$