AN ADAPTIVE, DYNAMICAL MODEL OF LINGUISTIC RHYTHM: DISSERTATION PROPOSAL

SEAN MCLENNAN

1. Introduction

Although adult native speakers of a language perceive spoken words as having clear boundaries and distinct components, the reality is that the speech signal rarely shows correlates (like pauses or silence) corresponding to those boundaries; it is primarily a continuous stream of modulated frequencies that is not trivially parsed into discrete linguistic units. The inability of scientists to find the physical primitives of speech despite more than fifty years of concentrated effort is at once a testament to the sophistication of the human auditory system and also an indication of our naive understanding of it.

It is perhaps easier to conceive of how adult speakers access the speech signal, given that they already have a well developed lexicon. The process could be seen as just a matter of matching the incoming signal to previously learned acoustic patterns; however, this is simply not a viable mechanism for infants who have yet to learn a single word. When could it be more important to be able to parse speech into linguistic units than during acquisition? We must conclude that even if acoustic pattern-matching plays a role in adult speech comprehension, it is just too simplistic a story to account for all of speech segmentation.

Date: April 6, 2004.

Time also proposes a significant problem. Physicists have a firm understanding of what time is, but it is clear this sense of objective time is not the "time" that is relevant to understanding biological phenomena. Even though, as speech researchers, we commonly measure and think in milliseconds, it may be that on a cognitive level, absolute time is less relevant than "linguistic time", although as we shall see, this too is something we do not have a firm understanding of.

Particularly for the task of segmenting speech, in the last couple of decades, significant focus has been directed at "rhythm" as the relevant measure of time mechanism for accessing speech. There seems to be a correlation between languages of different rhythmic types and strategies their speakers use to segment words and syllables. However, just as the human perception of discreteness in speech is deceiving, so too is our perception of rhythm. Historically, the underlying assumption has been that isochrony underlies rhythm—that the relevant rhythmic unit (syllable or stress, for instance) recurs at near regular intervals in speech. However, empirically, we have failed to find reliably isochronous intervals in natural speech that correspond to the accepted rhythm categories and we are forced to question the fundamental validity of those categories.

Evidence seems to be leaning towards the conclusion that the traditional categories are relevant and that the underlying reality behind our perception of rhythm is something more complex than simply isochrony. Rhythm remains implicated in a wide variety of cognitive functions, and is a compelling candidate for a linguistic bootstrap into

speech segmentation. What follows in sections 2 through 4 is a review of the research in linguistic rhythm, its impact segmentation, and how we might get a handle on what the nature of linguistic rhythm is. In section 5 I will propose a simple mechanism by which different rhythmic segmentation strategies may emerge based on experience with linguistic input.

2. Linguistic "Rhythm"

There is a long tradition of classifying languages according to human perception of their prosodic structure. James (1940), as an early example, refers to languages like Spanish and Italian as having a "machine-gun" rhythm. Pike (1945) later attributed the difference in the subjective rhythm of languages like English and Dutch from languages like Spanish and Italian as being "stress-timed" and "syllable-timed" respectively. A third category—"mora-timed"—to describe languages like Japanese in which timing seems based on a sub-syllabic unit (Abercrombie, 1967; Port et al., 1996) rounds out what has become the standard set of possible linguistic rhythm types.

This tri-partite classification is compelling; it seems to correspond with linguists' and speakers' intuitions about prosody. On the most superficial level, consider the native poetry types of English, French, and Japanese (the canonical representatives of each rhythm type). Japanese haiku and tanka have strict constraints not on the number of syllables in each line, but rather mora, which are assigned to not only light CV syllables, but also to geminate consonants, long vowels, and coda

nasals. Sonnets however, a historically French style, are strict with respect to the number of syllables. Lastly, English limericks are flexible with respect to syllables, as long as the characteristic stress pattern is obeyed. Although it is certainly true that English speakers can write haiku and Japanese speakers can write sonnets, there is an intuitive appropriateness to each style and its native language. So, while styles of verse cannot provide us with strong evidence of rhythm-types, it nonetheless reinforces the prevailing conception of how these languages differ.

What is meant by the term "rhythm" is somewhat more vague. The underlying presumption behind terms like "stress", "syllable", and "mora timing" is *isochrony* (Abercrombie, 1967); that is that the respective units in each language occur at roughly equal intervals, giving rise to a sense of periodicity. Empirically, however, this has been shown to be a gross oversimplification. For instance, in English interstress intervals are correlated with the number of intervening syllables, as well as their internal structure (Bolinger, 1965). This variability defies any claims of isochronous stresses in English. Wenk and Wioland (1982) failed to find isochronous syllables in French and Dauer (1983), comparing English, Thai, Spanish, Italian, and Greek, did not find that English interstress intervals were any more isochronous than the other (syllable-timed) languages. Japanese morae on the other hand are much closer to being isochronous; it appears that each mora contributes a near equal duration to the utterance (Port et al., 1987). "Rhythm", like many other perceived phonetic properties, is proving

to be difficult to define given the continuous and highly variable nature of articulation.

To complicate matters further, rhythm appears tied up with other language specific characteristics like syllable structure complexity and vowel reduction. That is, stress-timed languages tend to permit more complex syllable structures (particularly in stressed position) and in unstressed positions, vowels tend to be shortened and their quality shifted towards schwa. Syllable-timed languages tend to be more constrained, and mora-timed languages tend to be the most constrained (Dauer, 1983). Why this relationship should exist between the perception of time and structure is not immediately apparent, but it is identifiable. Indeed, Dauer (1987) proposes using characteristics like these as a more objective and meaningful method of typologically classifying rhythm types than the the standard "stress", "syllable", and "mora".

One advantage to Dauer's proposal is that it presents a more practical solution for dealing with ambiguous languages. Catalan, which has syllable structure and complexity like Spanish and which we would expect to be syllable-timed, also unexpectedly shows vowel reduction. Polish presents the opposite conundrum; the complexity of its syllable structure is typical of stress-timing, but it does not exhibit vowel reduction. Unsurprisingly, there is disagreement amongst linguists on how these two languages should be classified (Ramus et al., 1999).

Clearly what we call "rhythm" touches on an underlying reality for the impact of rhythmic-type is empirically well-founded. Even infants can distinguish between languages of different rhythm types (but not

the same) when utterances have been low-passed filtered, thereby retaining only prosodic information (Mehler et al., 1988). Metrical structure that is absent from natural speech immediately appears in simple repetition tasks (Cummins and Port, 1998), and speakers of languages belonging to different rhythm types show differences in their behavior on such tasks (Tajima and Port, 2003). Thus the question becomes what is the foundation of linguistic rhythm, if it is not what we would expect from the definition of the word "rhythm"? Franck Ramus and his colleagues have some compelling answers to that question which are the subject of the following section.

3. Potential Signal Correlates of Linguistic Rhythm

A number of different typologies, measures, and systems have been proposed to compensate for the apparent lack of isochrony (ex. Lehiste, 1977; Dauer, 1987; Cummins and Port, 1998; Cummins, 2002). Ramus et al. (1999) proposes a set of three signal correlates of linguistic rhythm that has garnered attention recently.

Very little experience is required for infants to be able to distinguish between languages of different rhythm types, long before the emergence of their own phonologies (Mehler et al., 1988). Thus it stands to reason rhythm is grounded in distinctions that infants are sensitive to. There is evidence that infants can identify vocalic periods in speech (Mehler et al., 1996) at the outset of acquisition, and it may even be that vowels and consonants are separably processed by the brain (Caramazza et al., 2000). Intuitively this makes sense: the distinction between vowels, which account for most of the energy in the signal, and consonants,

which contribute little to no energy, is the most salient perceptual distinction in oral language.

Ramus et al. (1999) consequently assume only this very rudimentary parsing of the signal and measured the durations of vocalic and consonantal intervals.[1] These were the basis for three simple statistics:

- %V: the proportion of the total duration of the signal that is vocalic.
- ΔV: the *variation* in the duration of vocalic intervals.
- ΔC: the *variation* in the duration of consonantal intervals.

Note that a perfectly isochronous signal would have a %V of 50%, and a ΔV and ΔC of 0. This would mean that the durations of vocalic and consonantal intervals are perfectly predictable, equal to each other, and occur at integer multiples. Thus, %V, ΔC, and ΔV can also be regarded as a rough measure of how strictly metrical the language is, at least with respect to these intervals.

These measures were taken on samples from 8 languages—Dutch, English, Polish (considered stress-timed); Catalan, French, Italian, Spanish (considered syllable-timed); and Japanese (considered mora-timed)—and plotted on three two-dimensional projections of the three-dimensional space defined by the factors. These plots appear in Figures (1) through (3).

Figure (1) is of primary interest. The distribution of languages on this plane maps well to the traditional stress-syllable-mora categories. Figures (2) and (3) show a similar clustering of languages, but Figure

[1] In measuring vocalic intervals, no distinction is made between consecutive vowels or long vowels, and off-glides (but not on-glides) are included.

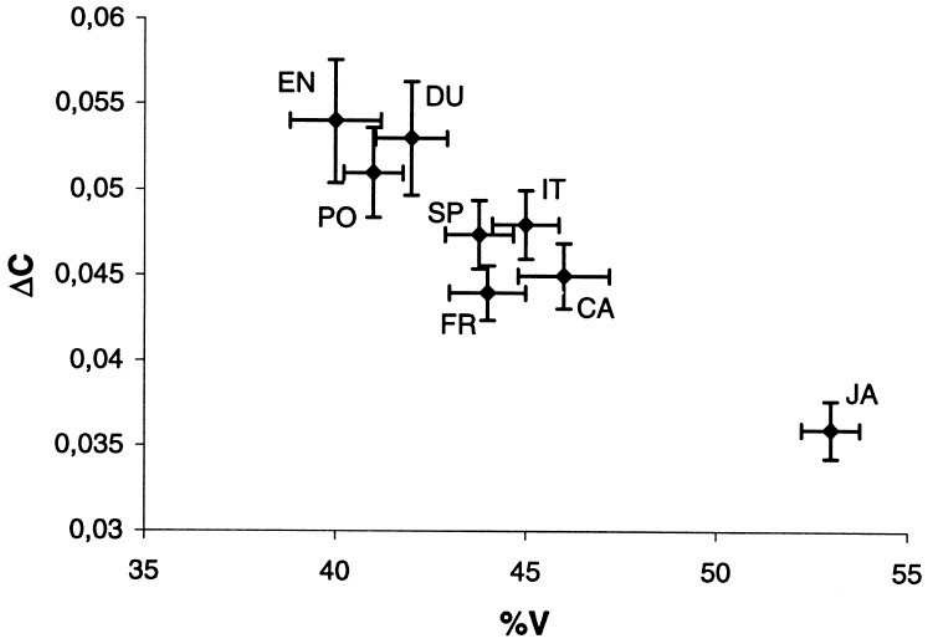

FIGURE 1. Distribution of languages over the (ΔC, %V) plane. Error bars represent ±1 standard error. (Ramus et al., 1999, pg. 273)

(1) shows a statistically significant effect of rhythm type whereas the others do not.

There are two important observations to make of (2) and (3), however. The first is that Polish is isolated from the other presumed stress-timed languages; this at once provides an explanation of the ambiguity concerning Polish's status in the literature, and suggests that Polish be treated as a separate category entirely. Indeed, further perceptual experimentation has shown that Polish is in fact distinguishable from English and Dutch (Ramus et al., 2003). This result further suggests that there may yet be other categories evidenced as more languages are examined with regard to these measures. The preponderance of

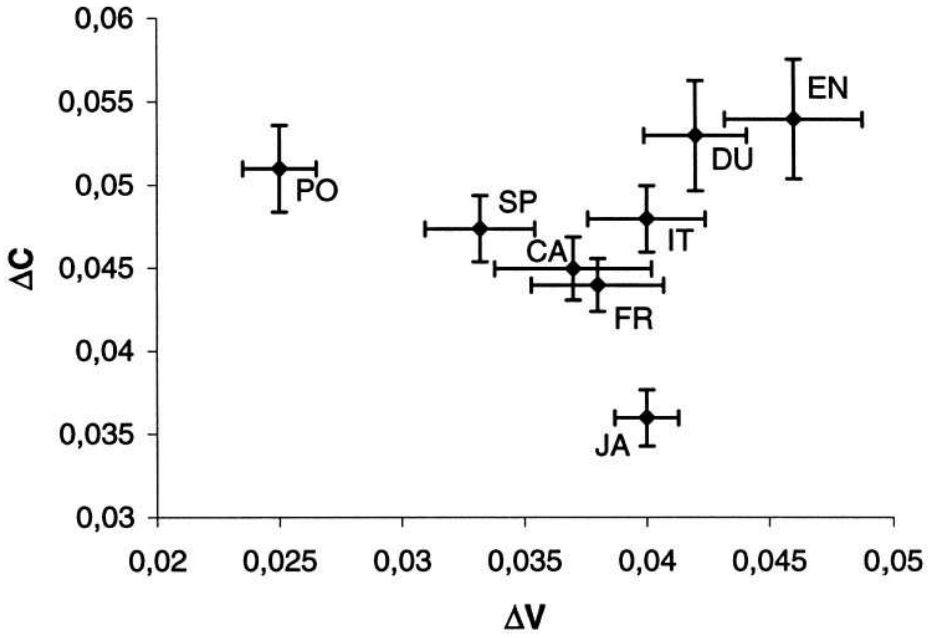

FIGURE 2. Distribution of languages over the (ΔV, ΔC) plane. Error bars represent $\pm$1 standard error. (Ramus et al., 1999, pg. 274)

the traditional three categories might then be due to markedness constraints as is similarly suggested by Levelt and van de Vijver (1998), or simply to an investigatory selection bias.

The second important observation is that although Japanese has a middling measure of ΔV, the error is relatively low—indeed the lowest of any of the languages. Ramus et al. (1999) did not provide the distributions of vocalic durations, but we can infer that the high variation and low standard deviation was probably caused by a bimodal distribution with values tightly clustered around two peaks, one of which is substantially higher than the other. This of course would reflect the fact that, in Japanese, vowel length is contrastive and that long vowels,

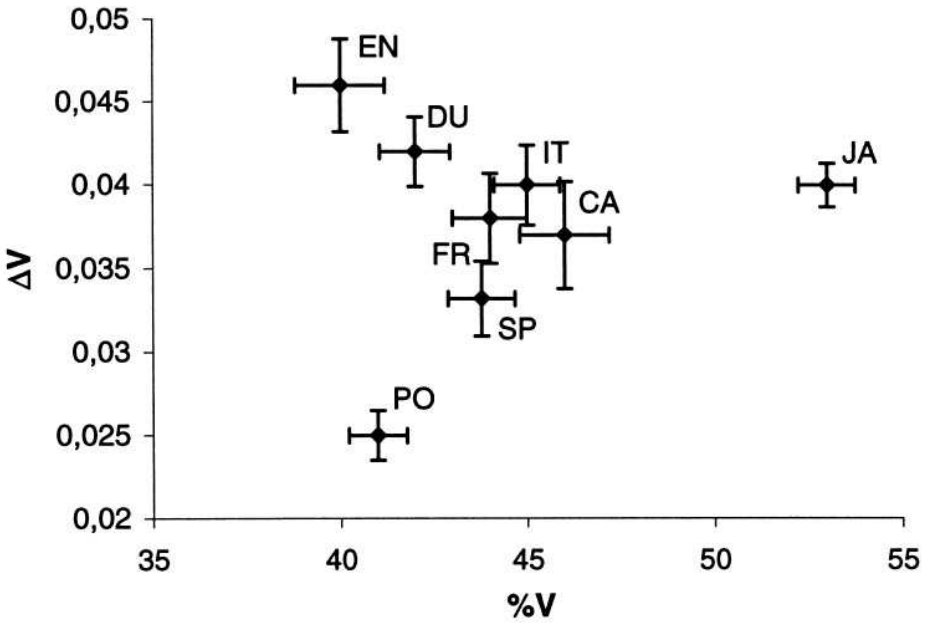

FIGURE 3. Distribution of languages over the (%V, ΔV) plane. Error bars represent ±1 standard error. (Ramus et al., 1999, pg. 273)

having two associated morae, are nearly exactly twice the length of a short vowel. This will be relevant to predictions of the model proposed here in section 5 below.

Ramus et al.'s result is compelling in the simplicity of the measurements, their grounding in the cognitive and linguistic development of infants, and in their ability to characterize the long standing categorization system that has proven itself relevant. It does leave open a number of important questions and %V, ΔC, and ΔV's suitability and reliability in characterizing cross-linguist rhythm type is the subject of ongoing debate (Ling et al., 2000; Barbosa, 2002; Cummins, 2002; Galves et al., 2002; Grabe and Low, 2002). One important question

is the interaction of speech rate, which although controlled for, is not varied in Ramus et al.'s study. Another is whether rhythm type is categorical or continuous in nature. As more and more diverse languages are analyzed according to these criteria, will they continue to cluster according to type, or will we find a continuous spread? Is the clustering seen in the present study simply the result of the selection of languages that prototypically represent their presupposed classes? Both of these questions are relevant to the research proposal outlined here.

At this point, we will momentarily set aside Ramus et al.'s measures and consider a quite different literature on the relationship of rhythm to speech-stream segmentation. A bridge between these two related areas is an aspect of this proposal that is addressed below in section 5.

4. Rhythm and Segmentation

Native speakers of a language effortlessly perceive the existence of discrete words in the speech stream, even perceive gaps between them, much like the spaces between printed words. In fact, such boundaries are largely absent from the signal. This poses the problem of how words are segmented from the speech stream.

A number of prominent recognition and lexical competition models have been proposed (McClelland and Elman, 1986; Norris, 1994; Luce and Pisoni, 1998) but these all rely on an existing lexicon for correct segmentation and recognition to occur. This provides little or no insight into how prelinguistic infants manage to perform initial segmentation. Although there may be some characteristic differences in the child-directed speech, its fundamental continuous nature is unaltered. It

may be that linguistic rhythm is the bootstrap that gives infants their first cues to segmentation (Cutler et al., 1992; Cutler and Mehler, 1993; Cutler, 1994).

Cutler and Carter (1987) observed that strong syllables (syllables containing a full vowel) in English provide a reliable segmentation cue; words in English are three times as likely to begin with a strong syllable as a weak syllable and words that begin with strong syllables occur with a higher frequency. Together, these facts estimate that about 85% of the word boundaries of lexical categories in English speech precede a strong syllable. Since strong syllables are related to prominence and greater perceptual salience, it stands to reason infants may use them as an initial heuristic. Similarly Mehler et al. (1981) found a potential "syllable effect" in French speakers.

These observations have led to more than two decades of research into the relationship between rhythm and segmentation cross-linguistically and there is good evidence that speakers of different languages segment speech using different criteria, and moreover that those criteria are grounded in the traditional "stress-timed", "syllable-time", or "mora-timed" classification system. (Cutler (1997) provides a general overview of the techniques and results.)

Experimentation has been centered on English, French, and Japanese—obviously because of their status as the "prototypical representatives" of their respective rhythm classes—and has used variations of a word-spotting task (McQueen, 1996) as the primary method of investigation. The task involves asking subjects to attend to a target word or syllable, and respond if and when they hear that target embedded within

longer words. Response times are measured, and the speed of response is taken to correlate with the relative ease of segmentation; that is, subjects respond slower if the target crosses expected segmentation boundaries.

The study by Mehler et al. (1981) and replicated in Cutler et al. (1986) that found a syllable effect in French, showed that French speakers will find a CV syllable faster in a CVCV- word than a CVCCV- word and a CVC syllable faster in CVCCV- word than a CVCV- word. For instance, if the target is *bal*, it is more readily spotted in *balcon* than *balance*. Conversely, if the target is *ba* it will be found faster in *balance*; subjects are significantly faster and finding targets that match the syllable structure of the stimuli.

This crossover effect in response times is completely absent in English speakers. Cutler et al. (1986) extended Mehler et al. (1981)'s finding by looking at both English and French speakers using both French and English targets and stimuli. Regardless of whether French speakers were listening to English or French, the effect appeared and regardless of whether English speakers were listening to English, French, or nonsense words, the effect was absent. English speakers show no preference for syllable structure in spotting target syllables.

Using an analogous task, however, Cutler and Norris (1988) showed that English speakers show a "stress-effect". Subjects can find the word *mint* faster in the nonsense word *mintef* than in *mintayf*. It is argued this is because the second strong syllable in *mintayf* induces a word segmentation boundary between [n] and [t] and finding *mint* thus requires reassembling information that has been already divided. They

are equally fast at finding *thin* in *thintayf* and *thintef*, a comparable task in which the target does not straddle the supposed boundary. Cutler and Butterfield (1992) provided further evidence for the effect by experimentally inducing missegmentations.

Naturally, given these results, we would expect to find an analogous "mora-effect" in Japanese and Otake et al. (1993) and McQueen et al. (2001) showed exactly that. Japanese speakers can find targets better when they correspond to mora boundaries than when they cross mora boundaries. Thus subjects can find the target *uni* in *gyan'uni* and *gyaouni* but completely fail to find it in *gyabuni*. Moreover, English and French speakers listening to the Japanese utterances perform appropriate to the previous results: French speakers show a characteristic crossover effect with respect to syllable structure that is absent from English speakers.

The results from English, French, and Japanese are convincing; they provide positive evidence in the debate over the relevance of the traditional rhythm classes. More detailed analyses of other languages needs to be performed before any conclusions are drawn, however; initial investigations are not as clear cut (Cutler, 1997) and it could be that the selection of these prototypical languages has overextended the generalizability of the results.

It is also clear that Cutler and her colleagues presume isochrony as the underlying tendency of rhythm, and indeed in Cutler et al. (1992) it is argued that infants respond to the lowest level of periodicity exhibited in their language. Given that this appears to be incorrect, where does that leave their findings? If Ramus et al. (1999)'s %V, ΔV, and

ΔC measurements are presumed to be the reality behind rhythm, how and why should variation in consonantal and vocalic intervals have the exhibited effect on segmentation? The line of research proposed here explores a hypothesis of how Ramus et al. and Cutler et al.'s findings interface.

5. Proposed Line of Research

The primary question this proposal addresses is "Can a naive adaptive dynamical model responding only to %V, ΔC, and ΔV produce behavior indicative of different rhythmic segmentation strategies?" If such a model is successful, it would provide support for Ramus et al's measures by providing an important bridge to Cutler et al's experimental results and it would also a powerful tool for exploring further experimental predictions.

In a Hidden Markov Model (HMM) speech recognition system, a window of fixed sized slides across a signal sampled in realtime measures (i.e. milliseconds) and it tries to match the input with statistical probabilities already learned. It seems unlikely that human speech perception occurs in this fashion. Humans don't perceive the signal in terms of milliseconds—they perceive it in terms of phrases, words, and segments that have an identity independent of the underlying realtime variations in their duration. Moreover, not all segments are created equal—vowels provide most of the energy in the signal and carry the most information. A more likely image than the HMM sliding window would be of a window *bouncing* down the signal not in fixed intervals measured in realtime, but being drawn to salient points in the signal,

effectively chunking speech into perceptual units and adapting appropriately. A system that progressed in this fashion would need to have two primary pieces of information: where those salient points are and how much of the signal the window should encompass. The "where" and "how much" of this proposed system are discussed below; it is however the "how much" that is most relevant to the question of how %V, ΔC, and ΔV might impact on segmentation.

5.1. **Where.** McLennan and Hockema (2002) describes a connectionist model that provides a biologically plausible method of implicitly measuring speaking rate from the waveform as it is being processed. The model itself is extremely simple, containing only two nodes, and is intended to be implemented in conjunction with another, more fully developed speech recognition system (the paper suggests the ARTPHONE model (Grossberg et al., 1997)). The output of the model is a spike train that marks roughly the centers of highly sonorant periods of the speech stream—i.e. the vowels. Sonorance is gauged using the intensity of the signal within the frequency band corresponding to the average human fundamental frequency (F0).

The most crucial aspect of the model is the dynamically changing weight from the input node to the output node. The adjusting of an input weight would usually be called "learning" in most neural network models, however in this case it is happening on a much shorter timescale, so it is referred to as "adaptation" in order to distinguish it from more permanent learning. This adaptation is the result of both Hebbian learning (Hebb, 1949) and habituation (Wang, 1995) employed simultaneously to adapt the same weight.

When the input activation (intensity of F0) is below a certain threshold (θ), the weight between the nodes is increased during periods of high intensity. However, once the activation of the neuron exceeds the threshold, it begins to habituate and the weight is gradually decreased. Thus, a key to the model is the balance struck between the increasing and decreasing portions of the weight dynamics during a period of sustained high intensity. At the end of such a burst, the net change in the weight will be a decrease if the burst was longer than expected and an increase if the burst was shorter than expected.

The dynamics here interact with the dynamics of another state variable, $A(t)$, the activation level. Thus, as the weight increases, this will cause the rate of change in the activation to increase. An increase in the neuron's activation will subsequently cause the rate of the weight adaptation to change (depending on if the neuron is above or below the θ threshold). Overall, this "snowballing" effect is useful in dealing with rapid changes in the speaking rate in real-time, and there is a normalizing effect on the output spike train. That is, the output spike trains for the same utterance spoken at different rates are very similar.

The dynamics are governed by the following equation:

$$\frac{dW}{dt} = \alpha * (\theta - A(t)) * I_{F0}(t) \tag{1}$$

where α is a learning rate parameter.

An interesting fact about the dynamics of this system is that if a vocalic period is encountered that is significantly longer than what is expected—that is significantly longer than the learned average—two spikes will be produced during the same vocalic period. This fact could

be very important for a language like Japanese that distinguishes between long and short vowels. As noted above in section 3, the distribution of Japanese vocalic intervals is likely strongly bimodal and favors short vowels—precisely the scenario that would give rise to multiple spikes.

The focus of Spike-V was employing habituation to provide a biologically plausible measure of speaking rate and consequently not much consideration was given to the exact point in the sonorous period that should be targeted and the center was arbitrarily chosen. It seems very likely however, that a point near the onset of the vowel—"P-centers" or "beats"—is more relevant for rhythmic tasks (Scott, 1993; Cummins, 1997). With small adjustments, Spike-V could probably be adapted to produce spikes close to these points, providing the "where" information required by this view of speech recognition

5.2. **How much.**

5.2.1. *The Window.* For recognition to occur, a single point does not provide sufficient information; some amount of prior and subsequent information from the signal needs to be integrated together. Hence, there must be a receptive "window" that delineates the portion of the signal being processed at that point in time. In this context, we usually think of "windowing" in terms of a segment of a waveform or spectrogram in which temporal information is available simultaneously. Really, the window is a spatial analogy for neural stimulus decay. Neurons respond to a transient acoustic event and it takes time for that excitation to

decay. Thus the size of the window is related to how long it takes a neuron to return to its resting activation.

A window can also be thought of as delineating information that will be chunked together into a single percept. By definition, portions of the signal outside the window are not available for consideration in the recognition of what falls inside the window[2]. In this sense, window boundaries are a good candidate for segmentation boundaries.

Consider the idealized signal in Figure (4). It is a simple graph of intensity and time where high intensity corresponds to a vocalic period and low intensity to a consonantal period. Thus, the graph represents a simple CVCV alternation in which each consonant and each vowel are identical in length—it is a perfectly periodic signal where %V is 50%, and ΔC and ΔV are 0. Because it is perfectly periodic, we know the exact size of a receptive window (the arcs above the signal) that would be required to minimally cover the entire signal.

However, consider Figure (5) which represents a CVCCVCV structure that has a %V that is less than 50%, a ΔV of 0, and a non-zero ΔC. Clearly the same fixed, narrow window will be inappropriate for recognition in this case as it will skip over important information about consonant clusters. We can state generally that the less periodic the signal, the wider the receptive window must be in order to ensure sufficient coverage. The more periodic the signal, the more one can afford to have a tight, fixed window size.

[2]Acknowledging of course that context and expectation can have a long-distance, top-down impact.

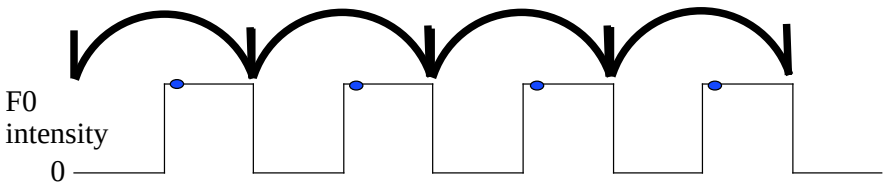

FIGURE 4. An intensity / time representation of an idealized signal. Absence of intensity corresponds to a consonantal period, presence of intensity corresponds to a vocalic period. The small dots represent a "where" spike as discussed in 5.1 above. The arcs above the graph represent the window.

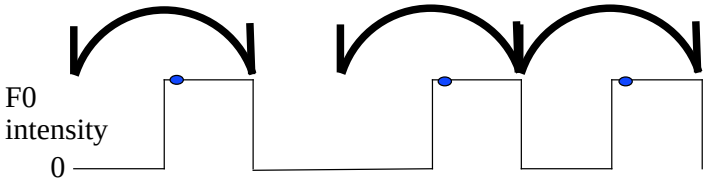

FIGURE 5. An intensity / time representation of a less than periodic signal.

The hypothesis that windows are related to segmentation boundaries and that window size is related to the relative periodicity of events in the speech signal (of which %V, ΔV, and ΔC are an estimate) suggests a mechanism by which rhythm class can impact on segmentation. It would be useful at this point to summarize a few observations about English, French, and Japanese (cf. Table 1).

The observations in Table 1 are consistent with the hypothesis that English speech recognition uses a relatively wide receptive window, Japanese a relatively narrow receptive window, and French, something in between. A wide window would be capable of encompassing more

	English	French	Japanese
Segmentation	stress	syllable	mora
Time Scale	supersyllabic	syllabic	subsyllabic
Periodicity	low	low	high
%V ΔV ΔC	low, high, high	mid, mid, mid	high, mid, low
Syllable Types	many	intermediate	few

TABLE 1

sequential segments, allowing for more complex syllable structures, but would likewise often include segments belonging to more that one syllable when those syllables happen to be simple[3]. A narrow window would severely constrain the complexity of syllable structure, and would be consistent with subsyllabic timing.

This is not to suggest necessarily that there is a causal relationship between the observations in Table 1, or perhaps more accurately, a direction of causality. If it is true that window-size is related segmentation strategies and syllable complexity, surely it is a bidirectional relationship and they are as intimately tied to each other as the chicken and the egg.

5.2.2. *Modelling the Window.* In Figures (4) and (5), there are fundamentally two acoustic events that provide temporal information: the onset of the vocalic interval and the offset of the vocalic interval. These two events are sufficient to determine %V, ΔV, and ΔC.

[3]This suggests that syllable boundaries may not always be clear, unintuitively, particularly when the syllable structure is simple; this is true in English

Suppose an adaptive oscillator (McAuley, 1995) was being driven by the onsets and offsets in Figure (4). It would very quickly entrain on a period equal to the consonantal / vocalic interval. If the oscillator were being driven by a more natural, less periodic signal, it's behavior would be more erratic. As it tried to adapt, expecting to find regularity, its period would fluctuate and generally it would fail to match events. If we think of the period as a prediction of where the next event will occur, we can think of the accuracy of that prediction as an estimate of how variable the interval is.

The picture is somewhat complicated by the fact that what needs to be estimated is not a single homogenous interval, but a single interval—suppose vowel onset to vowel onset—composed of two distinct subintervals: vowel onset to vowel offset and vowel offset to vowel onset. It is the error of those *subintervals* that needs to be gauged rather than the entire interval. While ΔV and ΔC seem somewhat correlated in Figure (2), Polish is an example where vowel intervals are considerably more predictable than consonantal intervals. Thus, ΔC, and ΔV do need to be estimated independently of each other.

This can be accomplished with a complex oscillator. Such an oscillator could be described in a number of different ways, but perhaps the most transparent visualization of it would be a point tracing a complex closed path such as the figure-eight in Figure (6). The diameter of each circle is related to the period of that subinterval; the sum is the period of the entire onset-to-onset cycle. Each diameter (period) can be adapted independently with the goal of each event, onset or offset,

coinciding with the trace at point A. The prediction accuracy of each period provides independent estimates of ΔC and ΔV.

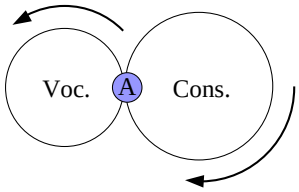

FIGURE 6. Visualization of the oscillator predicting vowel onset and offsets as a point following a figure-eight shaped path. Events are predicted to occur at point A.

The final measure, %V, is related to the difference in the diameter of the two circular paths in Figure (6). As the oscillator adapts over time, the diameter represents a learned estimate of the interval duration based on the intervals it has already encountered. Thus, the closer the two diameters are to being equal, the closer %V will be to 50%.

%V, ΔC, and ΔV can all be estimated from this one mechanism that adapts to the signal continuously as it is received. The estimates then can be used to adapt the size of the receptive window to a size appropriate to the language, and presumably, related to where language-specific segmentation boundaries occur.

6. Conclusion

6.1. **Time line.** I see this research project progressing on the following time line:

- Adapting Spike-V and assessment of performance: 2 months
- Development of the oscillator model: 1 month

- Replication of Ramus et al. (1999): 1 months[4]
- Assessment of performance on natural speech (consistent with above?): 2 months
- Assessment of potential segmentation based on window-size (consistent with Cutler et al.?): 2 months
- Exploration of unanticipated considerations; writing: 4 months.

6.2. **Discussion of Potential Results.** If this model is successful at producing behavior consistent with segmentation based on rhythm class, it will provide an important test of Ramus et al.'s hypothesis that rhythm class finds its perceptual basis in the three measures, %V, ΔV, and ΔC. However, the structure of the model, which distinguishes between "where" (the salient points in the signal relevant for rhythmic tasks) and "how much" (window size, hypothesized to be relevant for segmentation) also suggests that the concept of "rhythm" should be decomposed into quite distinct entities.

The model might also weigh in on the question of whether languages do cluster together into rhythm classes, or whether in reality they fall along a spectrum. Being a dynamical model, it is possible that attractor states may be apparent that correspond to rhythm classes. Also, because it is an automatic, computational model, it would facilitate the assessment of a wider diversity of languages.

Thus far, neither research into %V, ΔV, and ΔC, nor rhythmic segmentation has addressed the role of speaking rate which undoubtedly will impact on both. Because the model is rate-independent, it could

[4]Permission to use Ramus et al.'s corpus of recordings has been requested; pending response.

be used as a tool to explore differences in rate. Also, it suggests some interesting directions for such research; by necessity, as speaking rate increases, ΔC and ΔV will drop, narrowing the window-size. Thus it may be possible to manipulate segmentation experimentally by manipulating speaking rate based on predictions from the model. Moreover, highly rhythmic speech (poetry read aloud, speech cycling tasks, etc.) are more predictable by definition; it may be that there would similarly be observable differences in segmentations, particularly if syllable structure is tightly controlled to reduce variation in consonantal and vocalic periods.

References

Abercrombie, D. (1967). *Elements of general phonetics.* Aldine, Chicago, IL.

Barbosa, P. A. (2002). Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production. In *Proceedings of Speech Prosody 2002*, Aix en Provence.

Bolinger, D. (1965). *Pitch accent and sentence rhythm, Forms of English: Accent, morpheme, order.* Harvard University Press, Cambridge, MA.

Caramazza, A., Chialant, D., Capasso, R., and Miceli, G. (2000). Separable processing of consonants and vowels. *Nature*, 403:428–430.

Cummins, F. (1997). *Rhythmic Coordination in English Speech: An Experimental Study.* PhD thesis, Indiana University, Bloomington, IN.

Cummins, F. (2002). Speech rhythm and rhythmic taxonomy. In *Proceedings of Speech Prosody 2002*, pages 121–126, Aix en Provence, France.

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

Cutler, A. (1994). Segmentation problems, rhythmic solutions. *Lingua*, 92:81–104.

Cutler, A. (1997). The syllable's role in the segmentation of stress languages. *Language and Cognitive Processes*, 12(5/6):839–845.

Cutler, A. and Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31:218–236.

Cutler, A. and Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2:133–142.

Cutler, A. and Mehler, J. (1993). The periodicity bias. *Journal of Phonetics*, 21:103–108.

Cutler, A., Mehler, J., Norris, D., and Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25:385–400.

Cutler, A., Mehler, J., Norris, D., and Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology*, 24:381–410.

Cutler, A. and Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14(1):113–121.

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

Dauer, R. M. (1987). Phonetic and phonological components of language rhythm. *XIth International Congress of Phonetic Sciences*, pages 447–450.

Galves, A., Garcia, J., Duarte, D., and Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. In *Proceedings of Prosody 2002*, Aix en Provence.

Grabe, E. and Low, E. L. (2002). Durational variability in speech and rhythm class hypothesis. In Lahiri, A., editor, *Laboratory Phonology 7*, pages 515–546. Mouton de Gruyter, New York, NY.

Grossberg, S., Boardman, I., and Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance.*, 23:418–503.

Hebb, D. (1949). *The Organization of Behavior.* Wiley, New York, NY.

James, L. (1940). *Speech signals in telephony.* Sir Isaac Pitman and Sons, London.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, 5:253–263.

Levelt, C. and van de Vijver, R. (1998). Syllable types in cross-linguistic and developmental grammars. Paper presented at the Third Biannual Utrecht Phonology Workshop (11-12/06/1998).

Ling, L. E., Grabe, E., and Nolan, F. (2000). Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech*, 43(4):377–401.

Luce, P. A. and Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19:1–36.

McAuley, J. D. (1995). *Perception of Time as Phase: Toward an Adaptive-Oscillator Model ofRhythmic Pattern Processing*. PhD thesis, Indiana University.

McClelland, J. L. and Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, 18:1–86.

McLennan, S. and Hockema, S. (2002). Spike-v: An adapative mechanism for speech-rate independent timing. *Indiana University Working Papers Online*, 2.

McQueen, J. (1996). Word spotting. *Language and Cognitive Processes*, 11(6):695–699.

McQueen, J. M., Otake, T., and Cutler, A. (2001). Rhythmic cues and possible-word constraints in Japanese speech segmentation. *Journal of Memory and Language*, 45:103–132.

Mehler, J., Dommergues, J.-Y., Frauenfelder, U., and Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Behavior*, 20:298–305.

Mehler, J., Dupoux, E., Nazzi, T., and Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: the infant's viewpoint. In Morgan, J. L. and Demuth, K., editors, *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition.*, pages 101–116. Lawrence Erlbaum Associates, Mahwah, NJ.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., and Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29:143–178.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52:189–234.

Otake, T., Hatano, G., Cutler, A., and Mehler, J. (1993). Mora or syllable? speech segmentation in Japanese. *Journal of Memory and Language*, 32:258–278.

Pike, K. L. (1945). *The Intonation of American English*. University of Michigan Press, Ann Arbor, Michigan.

Port, R., Cummins, F., and Gasser, M. (1996). A dynamic approach to rhythm in language: Toward a temporal phonology. In *Proceedings of the Chicago Linguistic Society*. University of Chicago.

Port, R., Dalby, J., and O'Dell, M. (1987). Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America*, 81(5):1574–1585.

Ramus, F., Dupoux, E., and Mehler, J. (2003). The psychological reality of rhythm classes: perceptual studies. In *15th International Congress of Phonetic Sciences*, pages 337–342, Barcelona.

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

Scott, S. K. (1993). *P-centres in Speech: An Acoustic Analysis*. PhD thesis, University College London.

Tajima, K. and Port, R. (2003). Speech rhythm in English and Japanese. In Local, J., Ogden, R., and Temple, R., editors, *Phonetic Interpretation: Papers in Laboratory Phonology VI*, pages 317–334. Cambridge University Press, Cambridge, UK.

Wang, D. (1995). Habituation. In Arbib, M. A., editor, *The Handbook of Brain Theory and Neural Networks*, pages 441–444. MIT Press,

Cambridge, MA.

Wenk, B. and Wiolland, F. (1982). Is French really syllable-timed? *Journal of Phonetics*, pages 193–216.

Indiana University

E-mail address: mmclenna@indiana.edu