Linguistic Rhythm: A Dynamical Model

Sean McLennan
Indiana University
mmclenna@indiana.edu www.shaav.com

Overview:

It is well established that although adult speakers perceive spoken words as having clear boundaries, in reality the signal rarely shows correlates of those boundaries. While adult speakers might exploit a well developed lexicon to parse the speech stream, this is not a viable mechanism for infants who have yet to learn a single word.

Recent focus has been directed at "rhythm" as the relevant mechanism for accessing speech. There is a correlation between languages of different rhythmic types and strategies their speakers use to segment words and syllables (ex. Cutler, 1997). However, just as the human perception of discreteness in speech is deceiving, so too is our perception of rhythm. Rhythm implies an underlying isochrony which, empirically, we have failed to reliably find in natural speech.

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

My current research attempts to draw a bridge between two areas of rhythm research via a computational model: Ramus et al. (1999)'s views of what signal correlates underlie rhythm, and the impact of rhythm-class on the segmentation of words from the speech stream (Cutler, 1997). The primary question that is asked is "Can a simple learning mechanism (an adaptive oscillator model) that responds to Ramus et al.'s factors—the percent of the signal that is vocalic, the variance in the duration of vocalic intervals, and the variance in the duration of consonantal intervals—produce behaviour that is consistent with observed differences in segmentation behaviour?"

Preliminary results are reminicent of the patterns Ramus and colleagues have observed and ongoing research is promising. If ultimately successful, this model would provide support for the hypothesis that these factors underlie human perception of rhythm and would provide a plausible explanation for why these factors impact on how humans parse the speech signal, a heretofore unaddressed question. It could also give insight into an open question about the nature of linguistic rhythm: is it categorical or continuous?

Acknowledgements:

I would like to thank my dissertation committee: Michael Gasser, Linda Smith, Robert Port, and Ken de Jong for their mentorship, for sharing the depth of their knowledge, and for their sage wisdom.

I would also like to thank Franck Ramus for making available the data used by Ramus et al. (1999) which allows for much more meaningful comparisons.

Background:
The Reality and Perception of Rhythm

Rhythm Classes: There is a long tradition of classifying linguistic rhythm into three types: stress, syllable, and mora-based (exemplified by English, French, and Japanese respectively). It is a compelling system as it seems to correspond with the intuitions of linguists and speakers alike. Moreover, these classes seem to be perceptually relevant - for instance infants can discriminate languages of different classes, but not the same class (Mehler et al. 1988).

"Rhythm" itself, however, is an elusive thing to define. It implies an underlying isochrony that turns out to be absent, at least for syllable and stress-timed languages (Bolinger, 1965; Wenk and Wioland, 1982; Dauer, 1983). Morae on the other hand do seem quite isochronous (Port et al., 1987). To complicate matters, it seems that rhythm class is tied up with other linguistic phenomena like syllable structure complexity and vowel reduction and it isn't immediately apparent why there is a relationship between them. Indeed Dauer (1987) proposes that these characteristics are better used for typologically classifying languages than "rhythm".

Rhythm and Segmentation

While native speakers have little difficulty parsing an acoustic stream into a structure populated with discrete words, it is well established that there are no reliable signal correlates to word boundaries. This poses a particular issue for infants who, without experience or knowledge to rely on, nonetheless find words.

Rhythm has been implicated as a segmentation bootstrap in a large body of research conducted over the last two decades with infants and adults alike (Cutler, 1997 is a representative example). The primary insight that has emerged is that the traditional rhythm class of a language seems to have reliable impact on how native speakers segment words and syllables - not only in their own language, but when listening to languages of other rhythm classes as well. The reliability of these results supports "rhythm" as a valuable linguistic typological class.

However, again, this presumes isochrony as a fundamental, underlying tendency - indeed Cutler et al. (1992) aruge that infants respond to the lowest level of periodicity exhibited in their language. Given that the lack of isochrony seems as reliable a finding, it isn't clear where that leaves rhythm and segmentation.

Signal Correlates and a Different View of Rhythm

Ramus et al. (1999) proposes a set of three signal correlates of linguistic rhythm that have garnered attention recently:

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

The reasoning behind these measures is grounded in the observation that since naive infants to respond to rhythm distinctions, the signal correlates must be simple and salient; the relative difference in energy contributions of vowels and consonants to the speech signal make the contrast one of the most fundanmental linguistic distinctions.

Most importantly, when %V, ΔV, and ΔC measurements for a number of different languages are graphed against each other, the languages seem to cluster according to their rhythm classification. (Fig. 1-3). Eight Languages were investigated: English, Dutch, Polish ("stress-timed"); Catalan, French, Italian, Spanish ("syllable-timed"); and Japanese ("mora-timed"). Of particular interest is Fig. 2 where the clustering is most pronounced.

Thus far, these signal correlates are the most reflective of our expectations.

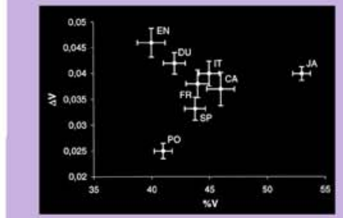


Figure 1: ΔV and %V
Ramus et al. (1999, 273). Interestingly Polish patterns quite differently than the other "stress-timed" languages. There is evidence that Polish should be considered a class all its own.

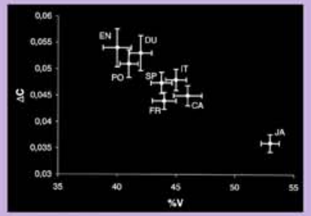


Figure 2: ΔC and %V
Ramus et al. (1999, 273). This projection shows the most pronounced clustering of the languages.

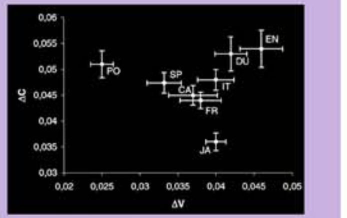


Figure 3: ΔC and ΔV
Ramus et al. (1999, 274). Again, Polish is exceptional. It is interesting to observe that Japanese has a middling of ΔV despite its apparent isocronous morae. This is likely reflective of Japanese' vowel-length contrast.

References:

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

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

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

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 Internaional Congress of Phonetic Sciences, pages 447-450.

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

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.

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., 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.

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

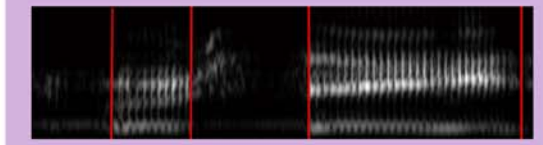
Bridging Signal and Segmentation:
An Adaptive Dynamical Model

There is reasonable evidence that %V, ΔV, and ΔC are good candidates for the underlying signal correlates of rhythm and likewise there is reasonable evidence that rhythm classes are relevant to segmentation. However, it is not immediately clear why there should be a relationship between %V, ΔV, and ΔC and segmentation (let alone some of the other rhythm-class correlated observations like syllable-structure and vowel reduction).

The present research proposes a three-part, biologically plausible learning mechanism that attempts to bridge the gap between the underlying reality and segmentation. Segmentation is conceived as an attentional "window" around salient points in the speech stream; the size of the window (correlated with rhythm-class) is driven by the variation in vocalic and consonantal intervals which is tracked by an oscillator attempting (and generally failing) to find periodicity in the signal.

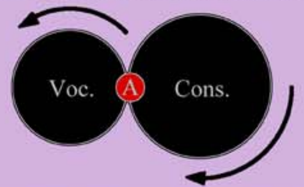
Tracking %V, ΔV, and ΔC

Two points in the speech signal need to be tracked regularly: vocalic onsets and offsets (which have been shown to be perceptually salient.



Suppose an oscillator were driven by these points. If the intervals between points were equal, ΔV and ΔC would be 0, %V would be 50%, and the oscillator would very quickly entrain on that periodicity. In a natural speech stream, however, where that periodicity does not exist, it would not likely entrain. It would however make *predictions* about where the next point would fall.

Since the oscillator would be responding to not to one homogenous interval, but two quite different intervals, its behaviour would be quite complex. There are multiple ways in which such an oscillator could be described but perhaps the most transparent visualization would be as a point tracing a figure-eight.



Encountering a vocalic onset, the trace would leave point A, traversing the left side of the figure-eight, and predicting the vocalic offset coinciding with its arrival back at point A. For the following consonantal period, it would traverse the other half of the figure. The diameters of each half are proportional to the learned / predicted duration of each interval - **the accuracy of the coincidence of point A with vocalic onsets and offsets is correlated with ΔV and ΔC. The difference in the diameter of each circle is proportional to %V.**

Figures 4 through 6 exhibit some preliminary results based on Ramus et al.'s (1999) data. Although not a qualitative replica of Figures 1 through 3, there are enough superficial similarities - particularly in 2 and 5 - to suggest that as the currently quite rough learning algorithm is refined, very similar patterns will emerge.

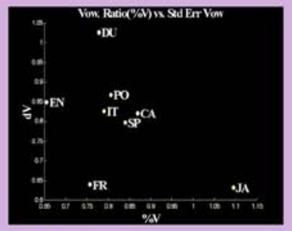


Figure 4: ΔV and %V
Results from the oscillator model using Ramus et al. (1999) data - cf. Fig 1.

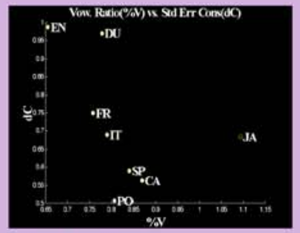


Figure 5: ΔV and %V
Results from the oscillator model using Ramus et al. (1999) data - cf. Fig 2.

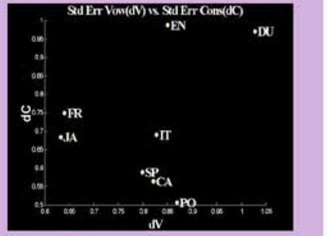


Figure 6: ΔV and %V
Results from the oscillator model using Ramus et al. (1999) data - cf. Fig 3.

Attentional Window

The bridge to segmentation in the proposed model is conceived as an "attentional window" centered on salient points in the signal. A single point in the speech stream does not provide sufficient information for recognition - there must be a "window" that delineates the portion of the signal to be processed at a given time. Unlike with a Fourier transform or an HMM, the attentional window in this context is really 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. In this sense, window boundaries are a good candidate for segmentation boundaries.

For a perfectly periodic signal, the window size would be easily defineable - it could equal exactly one consonantal-vocalic period and all information relevant to recognition would be guaranteed to be present. However, the more variation in vocalic and consonantal periods observed, the larger the window would need to be to ensure not losing important information. Thus, in this model, window size is taken to be directly proportional to ΔV and ΔC.

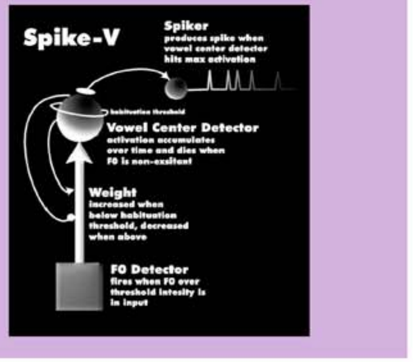
While this aspect of the model has not been yet developed, the hypothesis is that the window sizes will correlate with rhythm type, segmentation behaviour, and by implication syllable structure complexity and ambisyllabicity.

Finding Salient Points

In an HMM speech recognition system, a window of fixed size slides across a signal in realtime measures (ms). Humans don't perceive the signal in terms of realtime - they perceive it in terms of other semi-discrete timing units: segments, syllables, words, phrases etc. A more likely image would be of a window bouncing down the signal, being drawn to salient points in the signal, chunking speech into perceptual units.

The final piece of the proposed model is the "where"; what are the salient points in the speech stream? It seems likely that a point near the onset of the vowel - "P-centers" or "beats" which are implicated in a variety of rhythmic phenomena (Scott, 1993; Cummins, 1997) are good candidates.

McLennan and Hockema (2002) describe a connectionist model that provides a biologically plausible method of implicitly measuring speaking rate from the waveform as it is being processed. The model, Spike-V, (below) is extremely simple, consisting of only two nodes. Importantly, it uses both Hebbian learning and habituation to adjust the weight of its single connection which was effective at targeting the centers of vocalic periods. With small adjustments, Spike-V could be modified to target beats, providing the "where" aspect required by the model.


Spike-V
Spiker
produces spike when vowel center detector hits max activation
habituation threshold
Vowel Center Detector
activation accumulates over time and dies when F0 is non-existent
Weight
increased when below habituation threshold, decreased when above
F0 Detector
fires when F0 over threshold intensity is in input