A Neural Oscillator Model of Speech Timing and Rhythm

Erin Rusaw[1]

[1] Department of Linguistics, University of Illinois at Urbana-Champaign, USA

erusaw2@illinois.edu

Abstract

This paper introduces the Neural Oscillator Model of Speech Timing and Rhythm (NOMSTR), which is designed to be a flexible tool for investigating the systems which affect speech rhythm and timing through simulation. NOMSTR is an artificial neural network (ANN) model which incorporates oscillators, inspired by central pattern generators (CPGs), a type of neural circuit which underlies other types of patterned motor behavior in animals. NOMSTR uses three oscillators paired with thresholded nodes to model three levels of prosodic structure (e.g. syllables, accents, and phrases). In addition to setting the periods and phases of the oscillators to represent syllable and phrase durations, the weights between the thresholded nodes can be adjusted to model interactions between prosodic levels and their durational effects (e.g. pre-boundary lengthening). In this paper I demonstrate NOMSTR's ability to simulate the prosodic structure of spontaneous utterances in English and French, languages with disparate prosodic systems. The accuracy of NOMSTR's simulated prosodic structures is tested through its ability to simulate syllable durations, the locations of accents and phrase boundaries, and the influence of accenting and boundaries on syllable durations.

10.21437/Interspeech.2013-165

1. Introduction

An accurate model of part of the speech production system (in this case, the part of the speech planning system responsible for prosodic timing) must distinguish between those aspects of an utterance which are due to the properties of the speech production system, universal to all speakers, and which aspects are specific to an individual language. Thus, a successful model of speech production must have a core architecture that reflects the properties shared between languages, as well as variables that can be changed to simulate the differences between languages. Focusing on the temporal patterns of speech, a successful model must have an architecture that represents the elements that are related through temporal coordination (e.g.., gestures, syllables, feet, phrases), as well as parameters that can be varied to account for differences in timing patterns across languages.

Although the existence of truly regular time patterns in normal speech has been debunked (e.g. Dauer, 1983; Morton et al., 1976; Cummins, 2005), studies in multiple languages have demonstrated the usefulness of oscillator-based models in simulating suprasegmental timing in speech (e.g. Cummins & Port 1998, Saltzman et al. 2008, Barbosa 2002). One type of oscillator-based system which can provide a wider variety of interactions and more flexible timing in its output than mathematically coupled oscillators is the central pattern generator (CPG), which has been used to model other rhythmic motor behaviors (Delcomyn 1980). A CPG is a circuit of neurons in the central nervous system of an animal which generates a coordinated patterned output, or sometimes several different patterns of output. CPGs are commonly powered by sets of coupled oscillator neurons, which are responsible for generating their rhythmic output (Delcomyn, 1980). Speech production has much in common with other behaviors produced by or modeled with CPGs, such as locomotion: speech has underlying repetitive motions and an affordance for isochrony (Port & Tajima, 1999), flexibility in timing patterns (Tajima, 1998), adjustment to sensory and proprioceptive feedback (Gracco & Abbs, 1988), and the ability to integrate non-rhythmic motor tasks (Browman & Goldstein, 1989).

The Neural Oscillator Model of Speech Timing and Rhythm (NOMSTR), presented in this paper, is an artificial neural network model based on the CPG concept which is designed to simulate speech rhythm, or prosodic timing patterns. An earlier version of this model was previously used to simulate the durational effect on syllable duration of the interaction between accents and phrase-boundaries in English (Rusaw 2011). Here I use NOMSTR to model prosodic timing on a broader scale and with more naturalistic data, simulating the placement of accents and phrase boundaries and their effects on syllable durations across whole spontaneous utterances in both English and French.

2. Model Architecture

As mentioned in the introduction, a useful speech production model distinguishes between which aspects of the output (in this case, the suprasegmental timing of an utterance) are due to the nature of the universal production system, which are due to the language's prosodic system, and which are due to idiosyncrasies of the individual utterance. The universal architecture of NOMSTR which is the same for each utterance simulation, regardless of language, comprises three half-center oscillators which serve as the inputs to three thresholded integrate-and-fire artificial neurons, as shown in Figure 1.

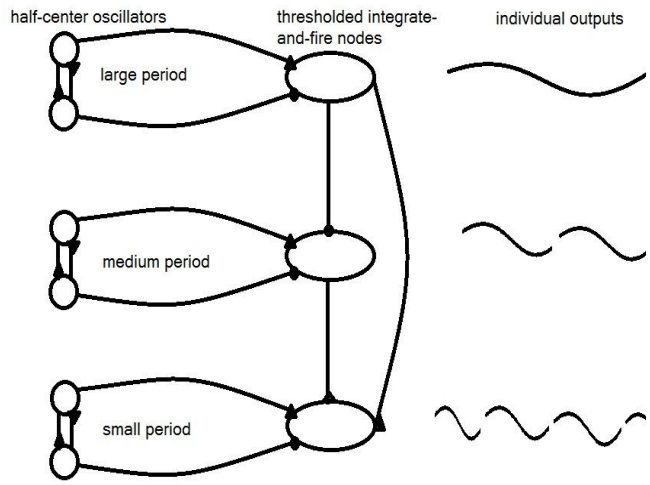

Figure 1: *Diagram of NOMSTR architecture*

One cell of each half-center provides excitatory input to its connected thresholded node, and one cell provides inhibitory input to the same. The resulting level of activation in each of the thresholded nodes is modeled as a sinusoid as shown in Equation 1. The three half-centers oscillate at three different frequencies, so that of the three half-center-thresholded node pairs, there is a small period, medium period, and large period set.

(1) activation = amplitude * sin(2*pi*frequency*((time) + phase));

Together these represent three metrical levels of phonological organization, such as syllable, stress group, and phrase; or syllable, accentual phrase, and intonational phrase. In this model, prominences (such as pitch-accents in English) and phrases are both modeled as cycles in the activation and firing of the artificial neurons. The duration between accents, for example, is measured as the duration between bursts of the thresholded node from the oscillator set representing the accent-level metrical tier; in the simulations reported here X these are generally the medium-period set (see also Rusaw 2011). Syllable duration and phrase length are measured as the duration between bursts of the thresholded node from the oscillator set representing the syllable (the small-period set) and the phrase (the large-period set), respectively.

The parts of NOMSTR which can be manipulated to model the different prosodic systems of different languages are the connections between the three thresholded nodes representing three levels of prosodic structure. All three thresholded nodes are connected to each other via excitatory or inhibitory connections, so that the total activation including inputs from other nodes can be described by Equation 2.

(2) total activation = (sinusoidal activation * weight of input from oscillator) + (output of node 1 * weight of connection 1) + (output of node 2 * weight of connection 2) + (output of node 3 * weight of connection 3)

If the activation level of a node does not reach the set threshold, it will not output, but if its activation level is above threshold it produces an output which reflects the sinusoidal curve of its oscillator. Excitatory or inhibitory influences on a node can have the effect of accelerating or delaying the next burst, respectively (and increasing or decreasing the simulated firing rate). As a result, the effect of one node firing can induce another node to fire, or accelerate or delay the next burst of another node, lengthening or shortening the duration of the node's current period. When used to model speech prosody, this can simulate prosodic events like the coincidence of a prominence with a phrase boundary, the coincidence of the boundaries of two phrases of different size; or the lengthening or shortening of syllable durations at prominences or phrase boundaries.

Finally, the speech rate and density of accents and phrase boundaries of a particular utterance are modeled in a NOMSTR simulation by adjusting the frequencies and phases of the three oscillators.

3. Simulating English and French Utterances

This study uses NOMSTR to simulate the prosodic structure of spontaneous utterances in English and French, focusing on the durational effects of accenting and phrase boundaries on syllables, and the location of affected syllables within each utterance. For the English simulations, the medium- and large-period (accent and phrase) thresholded nodes were set to inhibit the small-period node (syllable), and the large-period node was set to excite the medium-period node (for details see Rusaw 2011). For the French simulations, the medium- and large-period oscillators were set to inhibit the small-period node, and the large-period node was set to slightly inhibit the medium-period node. The English condition of this study uses three utterances from the English corpus described in Cole and Shattuck-Hufnagel (2011) for simulation. Each utterance was produced spontaneously during a map task in Shattuck-Hufnagel et al. (2004). The French condition uses three utterances from the Rhapsodie reference prosodic corpus of spoken French (Rhapsodie 2010), originally produced spontaneously during a direction-giving task. For each utterance, I recorded syllable durations (as measured from one c-center to the next (c.f. Browman and Goldstein 1988), and the locations of accented syllables and phrase boundaries, as identified in the corpora. For each utterance, these three measurements were used as the starting points for setting the oscillator parameters in NOMSTR, which were then slightly adjusted to optimize the simulation of the utterance.

4. Results

In the simulated English utterances, 100% of the syllables which touch (immediately precede or follow) a phrase boundary in the original production touche a phrase oscillator peak in the simulation. The simulations of the English utterances correctly predicted the locations of 94% of the accented syllables in the original productions (16/17), and 79% of the unaccented syllables (15/19). Although the French utterances contain slightly more syllables than the English utterances (14-15 vs. 12-13), each of the French utterances contains only a single IP, simulated by the model as a single peak of the IP thresholded node, which increases in output

across the syllables approaching the end of the phrase/utterance. The simulated French utterances correctly predicted the locations of 100% of the 10 AP boundaries in the French utterances, that is, each of the syllables which are AP-final the in original productions are covered by a peak of the AP oscillator in the simulations. The simulations also do not predict AP boundaries (peaks of the AP oscillator) on any syllables which are not AP-final in the original productions, although occasionally the beginning or end of an AP oscillator peak will "spill over" a small amount into the penultimate or post-boundary syllable.

The final phrase oscillator peak in the simulated utterance also covers the penultimate syllable, which of course does not directly "touch" a phrase boundary in the original production, but as studies of pre-boundary lengthening in English show (e.g. Shattuck-Hufnagel & Turk, 1998), could be influenced by the phrase boundary following the final syllable.

Figure 2 shows an illustration of English Utterance 1, as spontaneously produced by the speaker, and its simulation. Syllable durations are represented by the length of the red and blue bars, the locations of phrase boundaries in the original production are noted by black vertical lines, and accented syllables in the original are marked with asterisks. The locations of peaks of the accent nodes are marked by green horizontal lines, and the locations of peaks of the phrase node are marked by black horizontal lines.

Figure 2

As this figure shows, all of the syllables in S0's production that "touch" (immediately precede or follow) a phrase boundary are under the influence of the phrase oscillator peaks in the simulation. Additionally, in the simulation, the penultimate syllable of the phrase ("ra" in *paradise*) just touches the leading edge of the second phrase oscillator peak. Regarding accenting, all of the syllables in the original production which have pitch accents are under accent oscillator peaks in the simulation. In the simulation there is one syllable (*na*) under an accent oscillator peak which may not be accented in the real utterance; although *na* could be considered "stressed" as the metrical head of the word *Canadian*, the annotators of the speaker's production did not mark it as having an accent.

4.1. Syllable Durations and Lengthening

Figures 3 through 8 show for each of the three English and three French utterances: the durations of each syllable in the utterance for the original production and the duration of each syllable in the simulation.[1]

[1]The final syllables of the English utterances are not shown on the graphs, although they were simulated, because the corpus recordings were cut into individual utterances, preventing the c-center-to-following-c-center measurement for the final syllables.

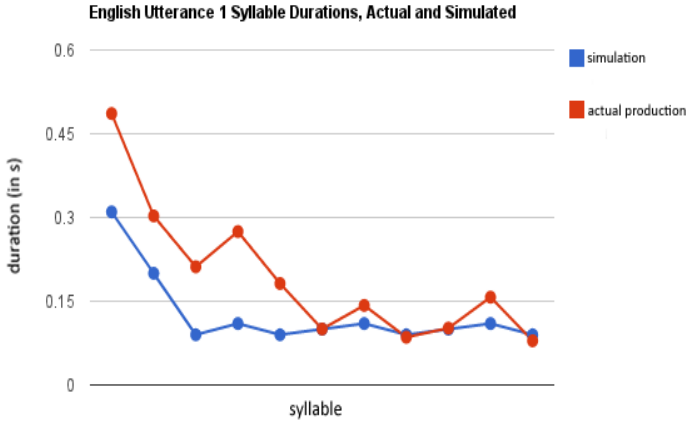

Figure 3: *English Utterance 1, Um Kate can you see the Canadian Paradise?*

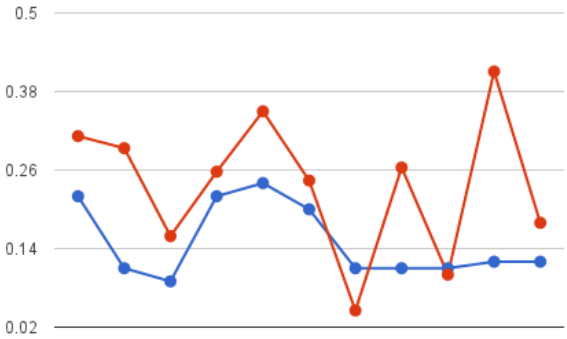

Figure 4: *English Utterance 2, And follow that path right above the fenced meadow*

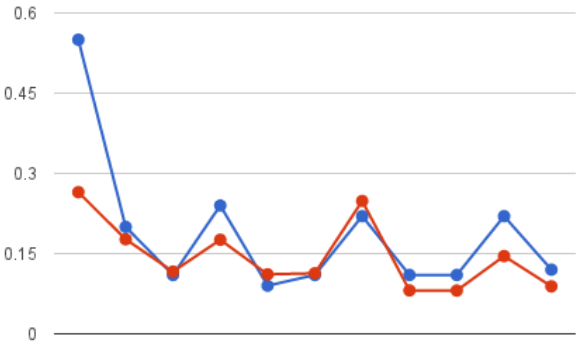

Figure 5: *English Utterance 3, No but you go to the left of the antelope*

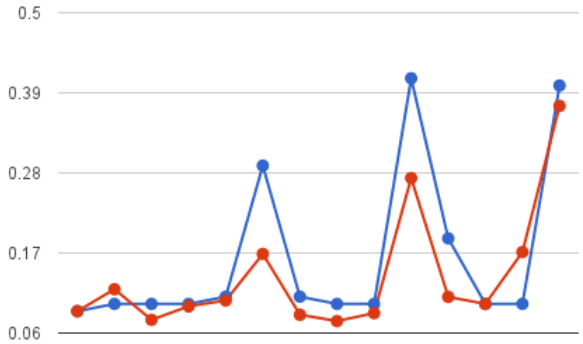

Figure 6: *French Utterance 1, Pendant un petit moment vous allez tojours aller tout droit*

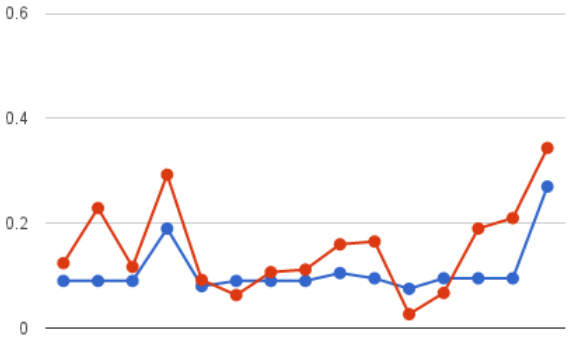

Figure 7: *French Utterance 2, Et le, le porche vous pouvez pas louper i le juste en face*

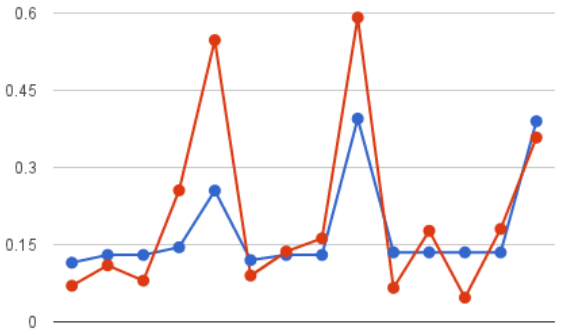

Figure 8: *French Utterance 3, Vous allez passer devant la poste qui sera a votre droit*

As these figures show, NOMSTR was successful in generating simulations of spontaneous utterances with a variety of different timing structures, in both French and English. This suggests that despite non-rhythmic influences on spontaneous speech, such as syllable content and idiosyncratic duration changes, much of the temporal structure of spontaneous speech can be modeled by a system whose timing rhythmic and *regular*, if not isochronous. NOMSTR is very accurate at simulating the locations of prosodic elements (syllables, accents, and phrases) within an utterance, as well as the influence these elements have on each other, most noticeably syllable lengthening. The notable exception is in English Utterance 2, for which the simulation failed to account for the lengthening in the 8th and 10th syllables, possibly due to inherent segment length (especially in the 10th syllable, *fenced*), or some other variable not accounted for by the model. Overall, the most consistent divergence between the utterances and simulations is that the simulations underestimate the amount of lengthening found in the longest syllables of either English or French utterances.

5. Discussion

The ability of the same basic system to model the prosodic timing structures of both languages provides some insight about the similarities and differences between the two prosody systems. First is the fact that at the level of timing and rhythm, prominences and phrases can be modeled in the same way; in NOMSTR, they are both cycles of an oscillator. The elimination of this distinction shows that the traditionally described difference between stress-accenting in English and accenting at the end of prosodic phrases in French is not necessarily a categorical difference at the level of the prosodic structure itself, but perhaps a difference in how the prosodic structure interfaces with the segmental, lexical, and syntactic information in an utterance; in what aspects of the "terrain" of an utterance (e.g. heavy syllables, syntactic phrase boundaries) are attractive anchor points for oscillation peaks. It is important to note that NOMSTR in its current form does not contain or account for any segmental information, which is of course a factor in the actual speech production system. Future integration of NOMSTR with information about the segmental content of the syllables in an utterance could lead to even more accurate simulations.

6. Future Work

There was one consistent issue with the simulated utterances: they tend to underestimate the degree of variation in syllable durations across an utterance. This could be somewhat accounted for by segmental content, for instance by very short reduced vowels in English, but it may also be due to the heavy influence that oscillators have on their paired thresholded nodes (which produce the output) in the current model, perhaps causing the syllable thresholded node to be less flexible in its timing than it needs to be. An updated version of NOMSTR which allows more elastic syllable durations is currently being tested.

7. Acknowledgements

This work was supported by NSF award 1155592 to Erin Rusaw and Jennifer Cole. Thanks to Drs. Jennifer Cole and Louis Goldstein for their comments

8. References

[1] Barbosa, P.A. 2002. Explaining Brazilian Portuguese resistance to stress shift with a coupled- oscillator model of speech rhythm production. Cadernos de Estudos Linguisticos. 43. 71-92.

[2] Browman, C.P. & Goldstein, L.M. 1988. Some notes on syllable structure in articulatory phonology. Phonetica. 45. 140-155.

[3] Browman, C.P. & Goldstein, L.M. 1989. Articulatory gestures as phonological units. Phonology. 6. 201-251.

[4] Buckley, E. 2009. Locality in metrical typology. Phonology. 26. 389-435.

[5] Cole, J. and Shattuck-Hufnagel, S. 2011 The phonology and phonetics of perceived prosody: What do listeners imitate? Proceedings of Interspeech 2011. Florence, Italy. 2011

[6] Cummins, F. & Port, R. 1998. Rhythmic constraints on stress timing in English. Journal of Phonetics. 26. 145-171.

[7] Dauer, R. 1983. Stress-timing and syllable-timing reanalyzed. Journal of Phonetics. 11. 51-62.

[8] Delcomyn F. 1980 Neural basis of rhythmic behavior in animals.Science.210.492–498.

[9] Edwards, J, & Beckman, M. E. 1988. Articulatory timing and the prosodic interpretation of syllable duration, Phonetica 45. 156-174.

[10] Edwards, J., Beckman, M. E., & Fletcher, J. 1991. The articulatory kinematics of final lengthening. JASA 89 1, 369-382.

[11] Grabe, E. & Low, E.L. 2002. Durational variability in speech and the rhythm class hypothesis. *Laboratory Phonology 7*. Gussenhoven, C. & Warner, N. eds. Mouton de Gruyter. Berlin.

[12] Gracco, V.L. & Abbs, J.H. 1988. Central patterning of speech movements. Experimental Brain Research. 71; 3. 515-526.

[13] Hayes, B. 1995. *Metrical Stress Theory: Principles and Case Studies*. The University of Chicago Press, Chicago.

[14] Morton, J.; Marcus, S. & Frankish, C. 1976. Perceptual centers P-centers. Psychological Review. 83;5. 405-408.

[15] Nespor, M. & Vogel, I. 2007. *Prosodic Phonology: With a new forward*. Mouton de Gruyter. Berlin.

[16] Pike, K. L. 1945. *The Intonation of American English*, Ann Arbor: University of Michigan Press.

[17] Port, R.F. & Tajima, K. 1999. Speech and rhythmic behavior. *The Non-Linear Analyses of Developmental Processes*. Savelsbergh, G.J.P.; van der Maas, H. & van Geert, P.C.L. eds. Royal Dutch Academy of Arts and Sciences. Amsterdam.

[18] Price, Patti J. / Wightman, C. W. / Ostendorf, Mari / Bear, John (1990): "The use of relative duration in syntactic disambiguation", In *ICSLP-1990*, 13-16.

[19] Rusaw, E. 2011. A biologically inspired neural network for modeling phrase-final lengthening. J. Acoust. Soc. Am. Volume 130, Issue 4, pp. 2553-2553.

[20] Rusaw, E. 2013. *Modeling Temporal Coordination in Speech Production Using an Artificial Central Pattern Generator Neural Network.* Unpublished Thesis, University of Illinois at Urbana-Champaign.

[21] Saltzman, E.; Nam, H.; Krivokapic, J; & Goldstein, L.2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. Proceedings of Speech Prosody 2008. Campinas, Brazil.

[22] Shattuck-Hufnagel, S., Dilley, L., Veilleux, N., Brugos, A. & Speer, R. 2004. F0 peaks and valleys aligned with non-prominent syllables can influence perceived prominence in adjacent syllables. In Proceedings of Speech Prosody 2004, 705-708, Nara, Japan

[23] Shattuck-Hufnagel, S. & Turk, A.E. 1998. The domain of phrase-final lengthening in English. Proceedings of the 16th International Congress on Acoustics. Seattle, USA. 1998

[24] Silverman, K.; Beckman, M.; Pitrelli, J.; Ostendorf, M.; Wightman, C.; Price, P.; Pierrehumbert, J. & Hirschberg, J. 1992. TOBI: A standard for labeling English prosody. Proceedings of the Second International Conference on Spoken Language Processing. Banff, Canada.

[25] Tajima, K. 1998. Speech rhythm in English and Japanese: Experiments in speech cycling.