

Generating Duration from a Cognitively Plausible Model of Rhythm Production

Plínio A. Barbosa

Lab. of Phonetics and Psycholinguistics & Dep. of Linguistics, IEL/UNICAMP, Brazil
plinio@iel.unicamp.br

Abstract

A dynamical model of rhythm production is presented. This model is meant to generate segmental duration from the interplay between a dynamical rhythmic system and a gestural score representation. The rhythmic level is being implemented by a coupled-oscillator system which delivers V-to-V size beats to the gestural score. To explain segment and pause acoustic durations, the interaction between rhythmic and gestural representations is achieved by a recurrent neural network. The model exhibits cognitively plausible language universal and language-specific phonetic properties which explain the variability of acoustic duration data.

1. Introduction

The dynamical approach to cognition [1] advocates that “natural cognitive systems are dynamical systems and are best understood from the perspective of dynamics” or, in classical terms, “to ignore movement is to ignore Nature” (Aristotle). Scientists “understand” a natural system when they are able to adequately model such a system, that is, when the variability exhibited at the model output closely follows the variability of natural data in response to equivalent input. Speaking or hearing subsume the interplay between two kinds of knowledge: linguistic representations (subject to the properties of formal syntax), and biomechanical and biochemical systems (subject to the laws of Dynamics).

Unfortunately, few models of speech production adequately capture these two kinds of knowledge, maybe because they involve the integration of discrete and real variables.

The two speech production models presented in the next section constitute an exception to that. On the other hand, recent technological solutions to speech synthesis ignore relevant issues in speech production modeling altogether.

2. Cognitively plausible models of speech production

Since the eighties, Articulatory Phonology [2], [3] (henceforth, AP) is a paradigmatic example of cognitively (and linguistically) plausible modeling. AP has straightforwardly explained phonetic and phonological variation from abstract, dynamical representations of articulatory gestures. These gestures are considered to be pre-linguistic forms of action. The success of this theory in modeling (linguistic) phonetic data is mainly due to the fact that abstract gestures have intrinsic time intervals which allow them to overlap in time. Some criticism against the theory has mainly concentrated on its alleged inability to taking into account categorical phenomena [4]. This kind of criticism can

be avoided by lexicalizing categorical allomorphy and allophony as well as by associating morphological labels to gestures’ edges [5]. To our understanding, a more serious drawback in AP framework concerns its inability to consider gesture coordination above the lexical level. In [6], a rhythmic tier is proposed to fill this gap, but the idea was not fully implemented and still lack the clear definition of extrinsic timing.

The so-called Temporal Phonology [7] explores this very notion and considers duration variability as a consequence of the dynamical coupling of a set of adaptative (or coupled) oscillators that are able to deal with metrics (structure) and cadence in speech rhythm.

Both Articulatory and Temporal Phonologies try to understand the nature and organization of phonological representations from accurate analyses of speech production data. As adequate models of speech production, both frameworks are able to generate acoustic parameters from abstract input and serve the goal of articulatory as well as acoustic speech synthesis research. Two other lines of research in speech generation systems can be considered as output-oriented models of speech production and are perhaps too close to market needs.

3. Output-oriented models of speech generation

Certainly in order to respond to technological demands, recent speech synthesis research preferred to assume its ignorance about the interplay between linguistic and physical parameters and to refuse to taking into account some well documented phonetic facts. Both use linguistic input as a way of classifying data to guide the record of huge speech corpora and to guide statistical analyses and searching procedures which accurately describe the acoustic output.

The so-called *corpus synthesis* [8] is an economically valuable technological solution to concatenative speech synthesis which refuses to investigate the interaction between prosody and segments as a worthwhile issue for explaining the variability exhibited by the acoustic parameters. This kind of work builds huge speech corpora containing units of variable size recorded under different prosodic conditions and uses complex search techniques which minimize paradigmatic discrepancy at the (abstract) input and which minimize syntagmatic discrepancy at the (physical) output.

The powerful statistical techniques used in van Santen’s work to model segmental duration [9] are an example of research blindness. By guiding its statistical analyses from some results on the relation of linguistic and phonetic variables, this kind of research is able to accurately predict segmental duration. But it fails in explaining crucial sources of variability as speech rate, in considering the problem of pause emergence (without postulating pause duration and location from the beginning of the generation process), and in taking into account some well documented phonetic facts. One of

these facts concerns durational phenomena above the segment level (the author is even “in complete disagreement with this kind of research” [9, p. 122]). The rhyme (and not the entire syllable) seems to be a universal prosodic unit [10]: phrase stress affects mainly this unit and marginally the onset consonants. Recently, we have shown that even across minor syntactic boundaries, both segments in the V#C unit are affected by phrasal stress degree [11] (stress groups end at phrasal stress). It is also mainly within the V-to-V unit (and not within phonological syllable limits) that compensation duration phenomena take place in several languages (such as BP, French and English, for which vowels are longer when followed by voiced fricatives and plosives which, on the other hand, are shorter than their voiceless counterparts). Considering V-to-V units (and not vowels or syllables) as stress bearers would explain, in a more elegant way than in [12], the duration *crescendo* of stressed syllables from proparoxytons to oxytons in Italian words, without the need of taking *ad hoc* decisions as considering vowel offset in oxytons at the onset of a following glottal plosive!

By trying to reproduce the output of the speech production mechanism from broad linguistic descriptions, these two output-oriented techniques serve only the needs of industry and cannot be considered as cognitively plausible models of speech production (and their advocates do not consider them as such).

Models that really take into account phonetic facts and take the risk of investigating complex phenomena like the role of prosody in modifying acoustic parameters at the segment level as well as the interplay between continuous and discrete variables in shaping speech output have more chance to succeed in explaining speech variability. Dynamical models of rhythm production belong to this class of challenging (but refreshing) research.

4. Coupled-oscillator models of rhythm production

The internal clock hypothesis says that cognitive systems have one oscillator that acts as a pacemaker which delivers regular beats to motor areas in the brain. This implies that all motor activities as breathing, walking, roaming, mastication and speaking would be timed by such a clock. Several authors have worked with this hypothesis (e.g., [13] and [14]).

Recently, some work on coupled oscillators have shown that coupled networks of neurons may generate oscillatory patterns of behavior in frequencies close to that of the mandible during speech (theta rhythms). These systems may exhibit several modes of oscillation that can explain the variability of complex activities as breathing and mastication [15]. In that way, coupled-oscillator systems are a more plausible version of the internal clock hypothesis.

In recent years, several authors have started exploring the notion of coupled oscillators for explaining subjects’ performance in speech production and perception (e.g., [1], [11], [16] and [17]) and even for explaining the variability of intergestural phasing in AP framework ([18]). O’Dell and Nieminen’s qualitative model has the advantage of directly referring to linguistic variables.

4.1. O’Dell and Nieminen’s qualitative model

O’Dell and Nieminen use a mathematical technique, Averaged Phase Difference (APD) theory, to obtain long-term,

qualitative descriptions of coupled oscillator models. By suggesting the coupling of a syllable and a stress group oscillator, they are able to explain the durational patterns of early research on speech isochrony based entirely on strict considerations of timing and strength of coupling (this coupling is achieved by a mechanism of entrainment by which an oscillator entrains a modification on the period of the other).

They demonstrate that the strength of coupling of the stress group oscillator on the syllable oscillator is in fact the ratio between the point of interception and the inclination of the regression line computed with the variables “duration of stress group” and “number of syllables within the stress group”. By doing so they are able to restate Dauer’s data [19] on isochrony in continuous terms: the higher the strength of coupling, the more stress-timed a language would be. They also show that the strength of coupling can vary with speech rate but they do not explore the consequences of this very fact, that is, two languages can only be compared with each other if both are spoken at similar speech rates. Even two varieties of a language cannot be considered as rhythmically equivalent without previous investigation [20].

5. A model integrating rhythmic and linguistic knowledge

Linguistic and rhythmic knowledge is being integrated in a straightforward way by a moment-to-moment dynamical model of rhythm production. This model is an improvement of an earlier version presented elsewhere [11]. The major improvement concerns the clear-cut separation of lexical stress representation at the gestural level from rhythmic properties as syllabicity and (phrasal) stressing.

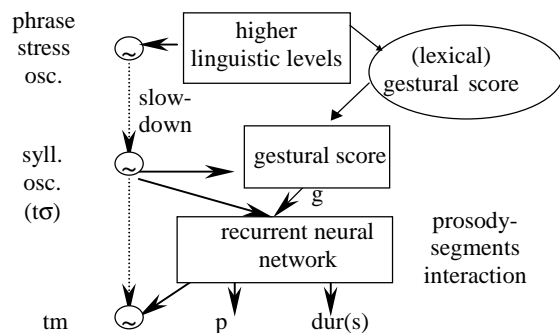


Figure 1: The dynamical model of rhythm production

The rhythm production part of this model concerns the coupled-oscillator system at the left, the insertion of the entrained syllabic oscillator beats at specific points in the gestural score, and the connectionist interaction between prosody and segments. In this framework, (phrase) stress and syllabic oscillators are considered universal properties of language because they are present as coupled oscillators in our cognitive system. The strength of coupling between the two oscillators is considered a language-specific property of the rhythmic system. Lexical stress is also considered a language-specific property and, according to [5], is the result of the difference in intergestural phasing between full and reduced gestures. The coordination of gestures within the gestural score follows the Acoustic-Articulatory Phonology developed by Albano from several studies on BP data [5].

Both oscillators are being implemented as trains of pulses with specification of period and magnitude. In Fig. 1, t_τ is the period of the abstract syllabic oscillator which is entrained by the (phrase) stress oscillator by a deceleration mechanism (slow-down). The effect of this mechanism is to increase V-to-V duration as phrasal stress approaches.

This phrasal stress is considered to be genuinely periodic but the exact location of its pulses is locally modified by higher-level linguistic input constraints (lexical and syntactic-semantic information).

The entrained syllabic oscillator acts as an extrinsic pacemaker for lexical gestural score: its beats determine the onset of the left edges of non-critical and non-closed Tongue Body gestures (vowel gestures). This gestural score *tout court* has then its V-to-V coordination given by the rhythmic system and the C-to-V coordination given by the rules of AP-like framework [5], [6]. This same abstract, cognitive, entrained syllabic oscillator entrains the natural oscillation period (t_m) of the mandible at the motor level.

Gestures (g) and entrained syllabic oscillator are the input of a connectionist recurrent network which delivers acoustic segmental duration (dur (s)), including silent pauses (p). The network can instead be trained to deliver duration of articulatory gestures and to serve the needs of articulatory speech synthesis research.

It must be clear by now that this model proposes to explain duration data variability as a consequence of both intrinsic (from intergestural phasing) and extrinsic (from the coupled oscillator model) timing. In this framework, segmental timing is a consequence of both macrorhythmic (within and above the syllable-size level) and microrhythmic (within and under the segment level) input.

Before completing the model and in order to evaluate if it will be able to generate segment and pause duration according to crucial sources of variability as speech rate, an improved version of the model already presented in [21] was evaluated. This version automatically generates acoustic duration for a BP TTS synthesis system [22].

6. First assessment: pause and acoustic duration generation in Aiuruetê

The model delivers acoustic segmental duration after two steps: connectionist V-to-V duration generation and statistical distribution of this duration among the segments. But the crucial information for succeeding network learning is to adequately correspond input variability to output variability.

As a modification to the earlier version, the network input contains two neurons representing syllabic (whose period specifies the speech rate) and phrase stress oscillators each. The period of the first oscillator is implemented as a real value in milliseconds. Following analyses from BP data, the second oscillator starts with a fixed period across speech rates (the phrase stress oscillator period). The ratio between this value and the period of the first oscillator is then computed and rounded. This new value represents the number of V-to-V units to the next phrasal stress. This number is readjusted in order to coincide with a lexical stressed syllable. For the time being, the only morphosyntactic information at the network input is the nature of some classes of functional words (separated by classes as determinants, conjunctions, prepositions and contractions of prepositions and articles). This higher-level information is considered to be sufficient to be used by

network to replace original phrase stress locations at periodically plausible boundaries.

A classical Elman recurrent network was implemented following the default specifications of MatLab®, version 6.0 (tansig activation function at the hidden layer) and the training (with gradient-descent algorithm) associated abstract input with normalized duration values of syllable-size units at the output. Duration data was obtained from the analysis of a corpus of 36 BP read sentences recorded at three speech rates.

As during speech production mechanism, the generation includes the emergence of silent pauses. As it was done for French [23], if at a particular position in the sentence (normally coinciding with phrasal stress beat) the corresponding delivered V-to-V normalized duration is greater than a critical value, the insertion of a silent pause is considered. If its duration is greater than a minimum (determined from analysis of the natural data) the procedure is over. If not, the high value of normalized duration naturally expresses the final lengthening of the corresponding segments within the V-to-V unit.

An example of acoustic segmental durations obtained is given in Table 1. The sentence (not used to train the network) is: “Ela coloca a sela do cavalo numa estante de uma antiga cela.” (she puts the (horse) saddle on a shelf of an old cell). The corresponding syllabic oscillator period specifying speech rate is also given.

Table 1: Generated acoustic durations (in milliseconds) for the sentence “Ela coloca a sela do cavalo numa estante de uma antiga cela.” The symbol “_” indicates a silence.

seg.	period of V-V osc.			seg.	period of V-V osc.		
	250 ms	200 ms	140 ms		250 ms	200 ms	140 ms
ε	134	101	85	n	56	32	29
l	56	48	43	ũ	133	115	112
ø	84	65	54	m	99	76	71
k	116	88	71	ø	98	43	38
o	140	102	85	e	127	78	58
l	76	51	41	s	124	33	14
o	160	125	111	t	73	81	37
k	115	86	70	ẽ	129	135	94
ø	92	75	62	t	86	80	46
a	158	74	48	ı	42	113	34
s	116	55	36	d	50	118	42
ε	173	160	151	i	97	89	70
l	96	85	123	ũ	112	121	100
ø	99	91	69	m	58	67	47
_	141	77	94	ø	48	71	43
d	128	112	72	ẽ	103	125	89
u	129	139	123	t	77	74	64
k	123	144	113	i	84	81	71
a	231	264	194	g	34	51	22
v	121	145	96	ø	63	69	49
a	228	193	214	s	201	191	166
l	87	107	81	ε	186	180	163
o	94	133	127	l	78	83	64
_	90	112	none	ø	144	148	129

It can be seen in Table 1 that the fastest speech rate generally presents shorter durations than the two others and do not place a silence after “cavalo”. The first slower rates differ mainly in the first half of the sentence.

7. Conclusions and Perspectives

The network is still being trained because the total mean square error is high and going down very slowly. It is probable that some variability of the input due to pitch accent placement needs to be considered as another input neuron in the network architecture. In this first evaluation, the coupling of both oscillators is being entirely done by the network. In the next steps, it will be dynamically achieved by the coupled-oscillator system itself from parameter specification of oscillatory behavior (see [16] for coupling in speech perception).

The ability of the network in selectively generating silent pauses according to speech rate is based only on the information at the input and is a major achievement of our model. The consideration of two levels of timing description within the coupled-oscillator system is associated with linguistic properties such as syllabicity and phrasal stressing. The choice of a V-to-V unit was determined by analysis on the acoustic data in French and in BP.

Due to its cognitive plausibility, such a model can authentically serve the needs of speech production modeling as well as articulatory and acoustic synthesis research.

8. Acknowledgments

We thank Eleonora Albano and Sandra Madureira for their helpful suggestions. This work is partially financed by a grant from a FAPESP project (n° 95/09708-6) and by a research grant (n° 350382/98-0) from CNPq, associated with the project n° 524110/96-4. It is also associated with the FAPESP project “Integrating Continuity and Discreteness in modeling Phonic and Lexical Knowledge”, n° 01/00136-2.

9. References

- [1] Port, R. F. and van Gelder, T. *Mind as Motion: Explorations in the Dynamics of Cognition*, The MIT Press, Cambridge, Mass., 1995.
- [2] Browman, C. and Goldstein, L. “Towards an Articulatory Phonology”, *Phonology Yearbook*, 3: 219-252, 1986.
- [3] Browman, C. and Goldstein, L. “Articulatory Gestures as Phonological Units”, *Phonology*, 6: 201-251, 1989.
- [4] Nolan, F. “The Devil is in the Detail”, *Proceedings of the XIVth International Congress of Phonetic Sciences*, August 1-7, San Francisco, USA, v. 1, 1-8, 1999.
- [5] Albano, E. C. *O Gesto e suas Bordas: Esboço de Fonologia Acústico-Articulatória do Português Brasileiro*, Mercado de Letras, Campinas, Brazil, 2001.
- [6] Browman, C. and Goldstein, L. “Tiers in Articulatory Phonology with Some Implications for Casual Speech”. In: Kingston, J. and Beckman, M.E. (Eds.) *Papers in Laboratory Phonology I*, Cambridge University Press, Cambridge, 341-376, 1990.
- [7] Port, R., Cummins, F. and Gasser, M. “A Dynamic Approach to Rhythm in Language: Toward a Temporal Phonology”, *Proceedings of the Chicago Linguistics Society*, Luka, B. and Need, B. (Eds.), 375-397, 1995.
- [8] Campbell, N. and Black, A. W. “Prosody and the Selection of Source units for Concatenative Synthesis”, In: van Santen, J.P.H., Sproat, R.W., Olive, J.P. and Hirschberg, J. (Eds.), *Progress in Speech Synthesis*, Springer-Verlag, New York, 279-292, 1997.
- [9] Sproat, R. (Ed.) *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer Academic Publishers, Boston, 1998.
- [10] Vaissière, J. “Language-independent prosodic features”. In: Cutler, A. and Ladd, D.R. (Eds.), *Prosody: models and measurements*, Springer-Verlag, Berlin, 53-66, 1983.
- [11] Barbosa, P.A. and Madureira, S. “Toward a Hierarchical Model of Rhythm Production: Evidence from Phrase Stress Domains in Brazilian Portuguese”, *Proceedings of the XIVth International Congress of Phonetic Sciences*, August 1-7, San Francisco, USA, v. 1, 297-300, 1999.
- [12] van Santen, J. and D’Imperio, M. “Positional Effects on Stressed Vowel Duration in Standard Italian”, *Proceedings of the XIVth International Congress of Phonetic Sciences*, August 1-7, San Francisco, USA, v. 1, 241-244, 1999.
- [13] Allen, G. D. “The Location of Rhythmic Stress Beats in English I & II”, *Language & Speech*, 15, 72-100, 179-195, 1972.
- [14] Turvey, M. T., Schmidt, R. C. and Rosenblum, L. D. “‘Clock’ and ‘Motor’ Components in Absolute Coordination of Rhythmic Movements”, *Haskins Laboratories Status Report on Speech Research SR-101/102*, 231-242, 1990.
- [15] Strogatz, S. and Stewart, I. “Oscillateurs Couplés et Synchronisation Biologique”, *Pour la Science*, n° 196, 40-46, 1994.
- [16] McAuley, J.D. *Perception of Time as Phase: Toward an Adaptive-Oscillator Model of Rhythmic Pattern Processing*. Unpublished PhD dissertation, Indiana University, USA, 1995.
- [17] O’Dell, M. and Nieminen, T. “Coupled Oscillator Model of Speech Rhythm”, *Proceedings of the XIVth International Congress of Phonetic Sciences*, August 1-7, San Francisco, USA, v. 2, 1075-1078, 1999.
- [18] Saltzman, E. and Byrd, D. “Dynamical Simulations of a Phase Window Model of Relative Timing”, *Proceedings of the XIVth International Congress of Phonetic Sciences*, 1-7 August, San Francisco, USA, v. 3, 2275-2278, 1999.
- [19] Dauer, R. M. “Stress-Timing and Syllable-Timing Re-Analysed”, *Journal of Phonetics*, 11: 51-62, 1983.
- [20] Barbosa, P. A. “‘Syllable-Timing in Brazilian Portuguese’: uma Crítica a Roy Major”, *D.E.L.T.A.*, 16 (2): 369-402, 2000.
- [21] Barbosa, P.A. “A Model of Segment (and Pause) Duration Generation for Brazilian Portuguese Text-to-Speech Synthesis”, *Proceedings of the Eurospeech’97*, Rhodes, Greece, v.2, 2655-2658, 1997.
- [22] Barbosa, P.A., Violaro, F., Albano, E.C., Simões, F.O., Aquino, P. A., Madureira, S. and Françoze, E. “Aiuuetê: a High-Quality Concatenative Text-to-Speech System for Brazilian Portuguese with Demisyllabic Analysis-Based Units and a Hierarchical Model of Rhythm Production”, *Proceedings of the Eurospeech’99*, Budapest, Hungary, September 5-9, v 5, 2059-2062, 1999.
- [23] Barbosa, P.A. and Bailly, G. “Generating Pauses within the z-score Model”. In: van Santen, J.P.H., Sproat, R.W., Olive, J.P. and Hirschberg, J. (Eds.), *Progress in Speech Synthesis*, Springer-Verlag, New York, 365-381, 1997.