Calculating articulatory syllable duration and prosodic boundaries

Donna Erickson[1, 3], Shigeto Kawahara[2], Jeff Moore[3], Caroline Menezes[4], Atsuo Suemitsu[5], Jangwon Kim[6], Yoshiho Shibuya[1]

[1]*Kanazawa Medical University, Japan*
[2]*The Keio Institute of Language and Cultural Studies, Japan,* [3]*Sophia University, Japan,* [4]*University of Toledo, USA,*
[5]*Japan Advanced Institute of Science and Technology, Japan,* [6]*University of Southern California, USA.*

EricksonDonna2000@gmail.com, kawahara@icl.keio.ac.jp,jeffmoore.personal@gmail.com, caroline.menezes@utoledo.edu, sue@jaist.ac.jp, jangwon@usc.edu, yshibuya@kanazawa-med.ac.jp

Abstract

Articulatory duration of vowels is often measured as the duration from jaw closure maximum to jaw closure maximum, flanking the production of the syllable nucleus. However, this method may not necessarily represent articulatory syllable duration, since the actual onset/offset of the syllable depends on the specific articulator (lip, tongue) that implements the articulation of the syllable onset/coda, and which does not strictly synchronize with timing of jaw movements nor with acoustically measured syllable durations (e.g., [Menezes, 2004]). We propose that by using a different approach, that suggested by the C/D model [Fujimura, 2000], it is possible to compute quantitatively the time values of prosodic boundaries from articulatory dynamics data. The algorithmic output, the "syllable pulse train", is the phonetic realization of the utterance's rhythmic structure (e.g., [Bonaventura and Fujimura, 2007]), which in turn reflects the phonologically derived metrical structure of the utterance (e.g., [Erickson et al., 2012]). Our small study presented here using the C/D model indeed revealed systematic articulatory patterns across speakers.

***Keywords**: C/D model, articulatory duration, prosodic boundaries, metrical structure*

1. Introduction

Acoustic syllable duration is traditionally measured as the temporal interval between variations in the acoustic waveform, e.g., [1]. However, it is possible that the articulation of the syllable may not be directly inferable from the acoustic output, e.g., [1, 2]. The articulatory duration of vowels can be measured as the duration from maximum jaw closure to maximum jaw closure, as the speaker articulates the nucleus of the syllable, e.g., [1, 3]. This approach may not always represent articulatory syllable duration accurately, since the onset/offset of the syllable depends on the specific articulator (lip, tongue) that implements the articulation of the syllable onset/coda. These movements do not strictly synchronize with the timing of jaw movements nor with acoustically identified syllable boundaries, e.g., [2].

We propose a different approach for defining articulatory syllable duration, based on the C/D model proposed by Fujimura [4]. Essentially, the model was proposed to account for the mismatch between acoustics and articulation, and was thus accordingly named, the Converter/Distributer model, in that it converts/distributes the phonological information to acoustic signals. The model is innovative in recognizing syllables of varying magnitudes as the basic structure of an utterance.[1] Syllable magnitudes are calculated based on jaw displacement for each syllable. A relatively prominent syllable has a larger jaw displacement than a relatively less prominent syllable. In the C/D model, an utterance is made up of a train of pulses that vary in height based on the syllables' magnitudes, and the syllable pulse height (syllable stress level) is commensurate with the articulatory syllable duration. This is interesting because recent studies show no consistent relationship between the syllable acoustic duration and stress level [2, 6]. Ongoing work by Erickson and colleagues suggests that the pattern of syllable pulse magnitudes within an utterance corresponds to the rhythm of the utterance, which in turn reflects the phonological metrical structure of the utterance, e.g., [7].

While the magnitude of the syllable is determined by maximum jaw displacement, the consonantal gestures of the syllable influence the timing of the syllable pulse.[2] According to the C/D model, the timing of the syllable pulse is centered relative to the speed (maximum and minimum velocity) of the crucial articulators of the onset and coda gestures. Crucial articulator (CA) refers to that articulator (tongue tip, tongue blade, tongue dorsum, lip) that articulates the onset and coda of the syllable. For example, the CA for [n] is tongue tip, for [p] is lower lip, and for [k] is tongue dorsum. Based on observation of an "iceberg" point (point with smallest mean invariance) in the overlaid demisyllabic velocity time function, the center of the syllable is defined as the midpoint between the syllable onset "iceberg" to the syllable coda "iceberg" [4,8,9]. However, in this paper we calculate the center of the syllable as the midpoint between the maximum speed of the crucial articulators following [10], described in more detail in Methods.

The C/D model describes not just syllable strength patterns, but also phrasing patterns [4]. If we know (i) the syllable strength (from the amount of jaw displacement), represented as various-sized syllable pulses, and (ii) the timing of the syllable pulse as described above, then we can calculate the articulatory duration of the syllable as well as the prosodic boundaries of the utterance.

In this paper, we use the C/D model to calculate prosodic boundaries and utterance rhythm from articulatory recordings of an English phrase, as spoken by four North American

[1] The intonation pattern (or melody) of an utterance is an aspect of utterance prosody, produced by laryngeal changes, described in part by its F0 contour, and is handled in the C/D model in a separate tier (see e.g., [5]). Therefore it is not addressed *per se* in this paper.

[2] There may be certain articulatory gestures, as in an interdental consonant, that prevent complete closure. This issue needs to be examined, but it is outside the scope of this paper.

English speakers. Previous studies have calculated prosodic boundaries from syllable triangles to examine a phrase with three monosyllabic digits [2, 9]. The current work uses the syllable triangle algorithm to examine the metrical structure of a four-word phrase: *nine tight night pipes.* The hypothesis is that by using the C/D model, systematic prosodic boundaries will be revealed across speakers, and these boundaries will be the same as those predicted by metrical theory. Moreover, their location (and possibly size) will match perception by listeners.

2. Method

2.1. Articulatory recordings and analysis

Acoustic and articulatory recordings were done using 3-D EMA (Carstens AG500 Electromagnetic Articulograph) at the Japan Advanced Institute of Science and Technology (JAIST), and at Haskins Laboratories, New Haven, Connecticut. One sensor was placed on the lower medial incisors to track jaw motion, one sensor each was placed on the tip of the tongue (TT), the mid of the tongue (TB) and the back of the tongue (TD). A sensor was placed on the lower lip (LL) and on the upper lip (UL). Four additional sensors (upper incisors, bridge of the nose, left and right mastoid processes behind the ears) were used as references to correct for head movement. The articulatory and acoustic data were digitized at sampling rates of 200 Hz and 16kHz, respectively. The occlusal plane was estimated using a biteplate with three additional sensors. In post processing, the articulatory data were rotated to the occlusal plane and corrected for head movement using the reference sensors after low-pass filtering at 20 Hz. The lowest vertical position (maximum displacement) of the jaw with respect to the bite plane was located for each target syllable of the utterance using the MATLAB-based custom software mview (Haskins Laboratories); this measure was used to indicate the height of each syllable pulse in the utterance. The position of the syllable pulse in the syllable was set at the midpoint between the maximum speed of the crucial articulator of the syllable onset and that of the syllable offset, also determined by a function in mview. The speakers were two male and two female North American English speakers. The utterance examined was, *Yes, I saw nine tight night pipes in the sky tonight,* adapted from [7,11]. Analysis is done for the phrase *nine tight night pipes,* which contains closed syllables with [aɪ] vowels. Before the data collection, the speakers had a chance to look at a picture illustrating the sentence, and could practice the sentence until they felt comfortable with it. The utterance was part of a larger corpus, presented to the speakers in randomized order, with five repetitions. The second or third utterance of each speaker is analyzed in this paper. The crucial articulators for the target syllables in this sentence are tongue tip (*nine, tight, night*), and lower lip (*pipe*). Figure 1 shows jaw displacement tracings for the phrase *nine tight night pipes,* the measurements of which are the syllable pulse height. It also shows displacement and velocity tracings for each of the Crucial Articulators.

Figure 2 shows the same utterance; the arrows mark the maximum speed of the Crucial Articulators of the syllables. Notice these are different from the time of maximum jaw displacement.

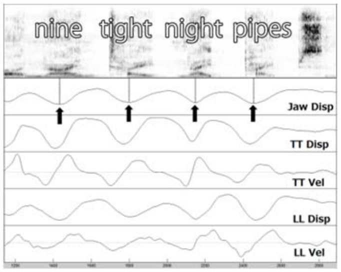

Figure 1: *"9 tight night pipes" from the utterance "Yes, I saw 9 tight night pipes in the sky tonight" as spoken by A3 (utterance 60). X- axis is time; y-axis is jaw displacement (mm) or articulatory velocity (mm/s). The spectrogram is shown in the top window, followed by Jaw Displacement, Tongue Tip Displacement, Tongue Tip Velocity, Lower Lip Displacement, and the bottom LL Velocity. The arrows point to maximum jaw displacement, the measurements of which are the height of the syllable pulse.*

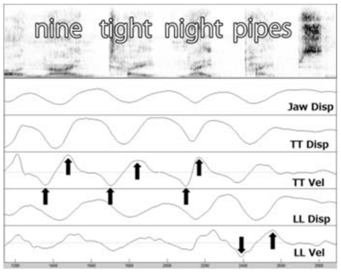

Figure 2: *Same utterance and captions as Fig. 1. Arrows mark the point of maximum and minimum velocity for each of the Crucial Articulators for the words, "nine tight night pipes".*

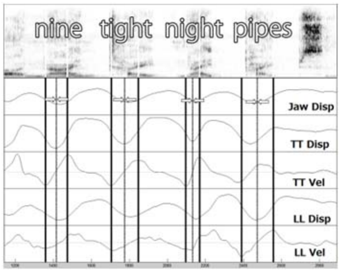

Figure 3: *Same utterance and captions as Figures 1 and 2. Arrows indicate the points midway between the maximum speed of the Crucial Articulators for each of the words, nine tight night pipes. Vertical lines indicate the points of speed of the onset and coda Crucial Articulators.*

Figure 4 shows how "syllable triangles" are constructed to represent each syllable. We calculate one constant angle, called "shadow" angle in the C/D model [4], for all triangles in an utterance in such a way that there is at least one pair of adjacent triangles whose edges meet and there is no overlap between any adjacent triangles. The length of the base of a triangle is the (abstract) syllable duration in the C/D model. The gap between two close edges of adjacent triangles is the duration of prosodic boundary between the two syllables of the triangles. We used the algorithm in the UBEDIT software [2] for computing theta for the isosceles triangles.

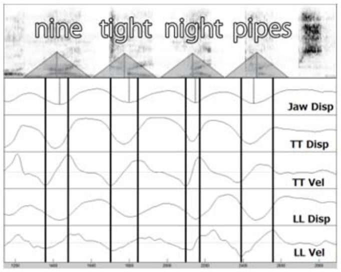

Figure 4: *Schematic illustration of syllable triangle construction for same utterance and captions in Figures 1, 2, and 3. Abstract syllable duration is the base of the triangle; apex is the jaw displacement/syllable magnitude.*

2.2. Metrical organization of the phrase

The metrical organization postulated for the four-word phrase is one major phrase (*nine tight night pipes*), and two recursive minor phrases *((tight)(night pipes))* [12,13], as indicated by the smooth brackets in Figure 5. According to metrical theory, one word in each phrase receives the largest stress; the major phrase stress can fall on either one of the minor phrase stresses. The minor phrase stress for this phrase will be on *nine* and on *pipes;* the major phrase stress will be on one and only one of these, i.e., *nine* or *pipes*, depending on the utterance conditions of the speaker. Here we suggest that *pipes* has the major phrase stress (based on work reported in [11]), and *nine*, the minor phrase stress. Numerical values of stress are arrived at by counting up the number of metrical grids (x's): the stress pattern hypothesized for this phrase is ((3)(1)(1 4)), where smooth brackets indicate phrasing, and, also shown in Figure 5. We assume that a boundary of a higher level is more likely to be realized, and will be longer in duration. Thus, the largest boundary will be after *pipes*, then after *night*, and then after *tight*. No boundary will appear in the two-word phrase, *night pipes*.

Major Phrase	(			x)
Minor Phrase	(x)	(		x)
Minor Phrase	x		(	x)
Word	x	x	x	x
Stress level	3	1	1	4
(Yes, I saw)	nine	tight	night	pipes

Figure 5: *Metrical grid for nine tight night pipes.*

2.3. Perception tests

A small-scale perception test was done with fifteen American college listeners for the four phrases analyzed in this study. Participants listened to the phrases and were instructed to mark phrasal structure, putting commas where small gaps/groupings are heard, following the methodology used by [14]. Before the actual testing, they were given a sample sentence, and asked to write commas where they heard a gap. They heard each utterance 8 times, in non-randomized order.

3. Results and discussion

3.1. Syllable pulse height and articulatory syllable duration

Table 1. *Syllable magnitudes (mm) and articulatory syllable durations (ms).*

	Syllable magnitude				Articulatory syllable duration			
Subj	nine	tight	night	pipes	nine	tight	night	pipes
A02	24.2	22.5	23.5	24.2	245.8	228.5	238.6	245.8
A03	23.7	23.3	22.7	23.9	266.5	262	255.3	268.7
A09	45.8	45.8	46.8	46.7	271.8	271.8	277.7	277.1
A10	23.1	21.6	22.4	23.5	305.5	285.7	296.3	310.8

Notice the positive relationship between syllable magnitude (syllable pulse height) and articulatory syllable duration. For instance, the syllable magnitude for A02 for both *nine* and *pipes* is 24.2 mm, the articulatory syllable duration for both is 245.8 ms. A positive relationship between syllable magnitude and boundary duration is interesting, especially in view of the finding by [6] of no significant correlation between acoustic syllable duration and jaw displacement. Moreover, using articulatory information as a means for determining syllable duration has advantages since sometimes it is impossible to measure duration from acoustic signals, as for instance, one does not know exactly where [t] for *night* ends and the initial [p] for *pipes* begins. Also notice that three of the four speakers show the largest amount of jaw opening for *pipes*. In the next section are shown the syllable triangles, followed with a discussion of the metrical structure and phrasing.

3.2. Syllable Triangles

Figures 6~9 show the results of applying the algorithmically-objective method for deriving articulatory syllable duration and prosodic boundaries, for the phrase *nine tight night pipes*, as spoken by four North American speakers. In the figures below, the height of each triangle represents the amount of jaw displacement/the amount of syllable magnitude (stress), the base of each triangle is the articulatory syllable duration, and the spaces between each triangle show the prosodic boundaries.

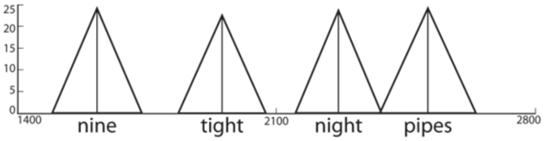

Figure 6: *Nine tight night pipes Speaker A02 (ut 33). The x-axis is time (ms), the y-axis, jaw displacement (mm).*

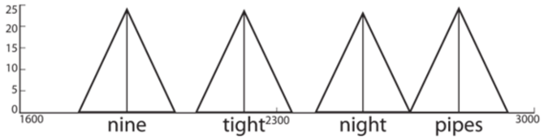

Figure 7: *Nine tight night pipes Speaker A03 (ut 60). The x-axis is time (ms), the y-axis, jaw displacement (mm).*

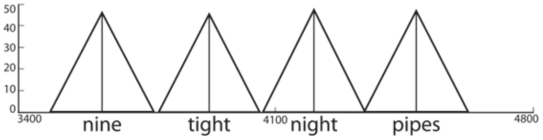

Figure 8: *Nine tight night pipes Speaker A09 (ut 19). The x-axis is time (ms), the y-axis, jaw displacement (mm).*

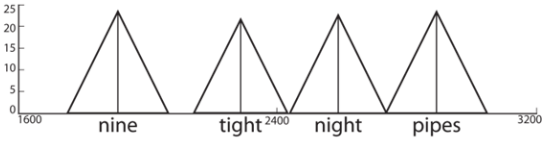

Figure 9: *Nine tight night pipes Speaker A10 (ut 113) .The x-axis is time (ms), the y-axis, jaw displacement (mm).*

3.3. Syllable triangles and metrical organization

In terms of syllable triangle height, the figures (also see Table 1) show that for three of the four speakers, *pipes* has the largest amount of jaw displacement, followed by *nine*. This matches the metrical structure depicted in Figure 5.

In terms of prosodic boundaries, as was predicted by Figure 5, we see for all speakers, two boundaries, one after *nine* and one after *tight*, but none between *night* and *pipes*. Moreover, as predicted by Figure 5, three of the four speakers show a bigger boundary after *nine* than after *tight*, as shown by the durations in Table 2 (left-most data columns).

Table 2. *Prosodic boundary durations (left-most 2 data columns) and perceived boundaries (right-most 2 columns). The numbers in bold type indicate the larger of the two boundaries*

Speaker	Measured boundary after nine	Measured boundary after tight	Perceived boundary after *nine*	Perceived boundary after *tight*
A02	**99.3 ms**	81.5 ms	53%	47%
A03	55.2 ms	**65.8 ms**	57%	43%
A09	**19 ms**	12 ms	55%	41%
A10	**80.4 ms**	10.6 ms	39%	61%

3.3. Comparison of syllable triangle boundaries with listeners' perceptions

Table 2 also shows the perceived boundary strengths, calculated in percentages as the number of commas placed at a given juncture compared to the total number of commas for a given phrase. For Speaker A09, the total percentage does not add up to 100%, because some listeners put commas after the final word, *pipes*, even though that was not the assigned task. The results of the very preliminary tests showed that listeners perception of location of boundaries agreed with the algorithmically calculated boundaries, that is, they heard breaks after *nine* and *tight,* and not between *night* and *pipes.* However, the strength of listeners' perceptions matched the numerical calculated boundaries for only two of the speakers. Perhaps this is because the perceptual task of evaluating pauses was a difficult one. Listeners often varied their placement of commas when presented with the same utterance eight times in a row even though they knew it was the same utterance. There are a number of reasons for the difficulty in perceptually evaluating strength of prosodic boundaries, including the fact that notation of boundary strengths of non-major phrases is not part of our writing system. Or it may be that a larger number of listeners are needed to elucidate minor prosodic boundaries, as was done in [15]. Listeners' perception of boundary strengths is part of our on-going investigation into the efficacy of the C/D model for understanding utterance prosody.

4. Conclusion

This paper is a preliminary study to apply the C/D model to calculate articulatory syllable duration and prosodic boundaries. The results of this small study suggest that analyzing articulatory data within the framework of the C/D model can provide prosodic information about stress and boundaries. Moreover, this approach is algorithmically very objective, and as such, produces replicable results. It is hoped that the type of information generated by calculating articulatory syllable duration and prosodic boundaries, using the C/D model, will lead to future, more comprehensive and objective studies of metrical structure and phrasing.

5. Acknowledgements

The impetus for this approach to the phonology-phonetics interface of language and its metrical structure comes from the insights of Osamu Fujimura. This work was supported by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (C)#22520412 and (C)#25370444. A special acknowledgement to Mark Tiede for help with mview, to Bryan Pardo for UBEDIT, and to Mark and to Ian Wilson, for participating in the experiment. Also, we thank Sungbok Lee for discussions about this paper.

6. References

[1] Edwards, J. and M. E. Beckman (1988). "Articulatory timing and the prosodic interpretation of syllable duration". In: *Phonetica*,45, pp. 156-174.

[2] Menezes, C. (2004). "Changes in phrasing in semi-spontaneous emotional speech: Articulatory evidences". *J. Phonetic Soc. Japan*, 8, pp. 45-59.

[3] Edwards, J., M. E. Beckman, and J. Fletcher (1991). "The articulatory kinematics of final lengthening". In: *J. Acoust. Soc. Am.* 89, pp. 369-382.

[4] Fujimura, O. (2000). "The C/D model and prosodic control of articulatory behavior". In: *Phonetica* 57, pp. 128-138.

[5] Fujimura, O., and D. Erickson, D. (2004). "The C/D model for prosodic representation of expressive speech in English". In: *Acoust. Soc. of Jp, fall meeting,* p. 271-2.

[6] Erickson, D., Y. Shibuya, A. Suemitsu, and M. Tiede (submitted). "Articulation of phrasal stress: A preliminary comparison of American and Japanese speakers of English".

[7] Erickson, D., A. Suemitsu, Y. Shibuya, and M. Tiede (2012). "Metrical structure and production of English rhythm". In: *Phonetica* 69, pp. 180–190.

[8] Fujimura, O. (1986). "Relative invariance of articulatory movements: An iceberg model". In: J. Perkell and D. H. Klatt (eds), *Invariance and variability in speech processes*, pp. 226-242, Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

[9] Bonaventura, P. and O. Fujimura (2007). "Articulatory movements and prosodic boundaries". In: P. Beddor, J. Ohala and M. Solé (eds.), *Experimental Approaches to Phonology*, pp. 209-227 Oxford: Oxford University Press.

[10] Erickson, D. (2010). "More about jaw, rhythm and metrical structure". In: *Acoust. Soc. of Jp, fall meeting,* p. 103.

[11] Erickson, D., S. Kawahara, J. C. Williams, J. Moore, A. Suemitsu, and Y. Shibuya, (2014). "Metrical structure and jaw displacement: An exploration". In: *Speech Prosody 2014* (Dublin, Ireland, May 2014).

[12] Selkirk, E. (2011) "The syntax-phonology interface". In: J. A. Goldsmith, J. Riggle, and A. C. L. Yu (eds), *The Handbook of Phonological Theory*, pp. 435-484. Second Edition, Hoboken, NJ: Wiley-Blackwell..

[13] Ito, J., and A. Mester, (2012). "Recursive prosodic phrasing in Japanese". In: T. Borowsky, S. Kawahara, M. Sugahara, and T. Shinya (eds.), *Prosody Matters. Essays in Honor of Elisabeth Selkirk. Advances in Optimality Theory Series. Elsevier.* Earlier version in the *Proceedings of the 18th Japanese/Korean Conference,* 2010. pp. 147-164. Stanford, CA: CSLI.

[14] Menezes, C., D. Erickson, J. McGory, B. Pardo, and O. Fujimura (2002). "An articulatory and perceptual study of phrasing". *Temporal Integration in the Perception of Speech. ISCA Workshop. (Aix-en-Provence, April 8-10)*, p. 43.

[15] Cole, J., L. Goldstein, A. Katsika, Y. Mo, E. Nava, and M. Tiede (2008). "Perceived prosody: Phonetic bases of prominence and boundaries". In: *J. Acoust. Soc. Am.* 124, p. 2496.