音声研究　第 19 巻第 2 号
2015（平成 27）年 8 月
16–21 頁

Journal of the Phonetic Society
of Japan, Vol. 19 No. 2
August 2015, pp. 16–21

特集書評

A Practical Guide to Calculating Syllable Prominence, Timing and Boundaries in the C/D Model[1)]

Donna Erickson* and Shigeto Kawahara**

C/D モデルにおける音節卓立，タイミング，境界計算の実用的指針

SUMMARY: The Converter/Distributor (C/D) model (Fujimura 2000) provides a comprehensive and explicit framework to model how the phonological, prosodic organization is mapped onto actual speech production. The goal of this paper is (i) to walk readers through how to construct the prosodic representations of the C/D model from actual articulation data, and (ii) discuss some crucial concepts of the C/D model. Some basic hypotheses of the C/D model are (1) phonological syllable magnitude increases with increased sentence stress, (2) amount of jaw displacement is the articulatory correlate of syllable magnitude, (3) phonological syllable timing is calculated from speed patterns of the crucial articulators of onset and coda consonants, and (4) once syllable magnitude and syllable timing are determined, we can automatically calculate phonological phrasing patterns, with phrase boundaries which come with predicted durational values. All of these computational aspects of the C/D model can and should be tested empirically. In this paper, we attempt to explain and discuss these aspects of the C/D model in detail, especially for those readers who are not already familiar with the model.

Key words: C/D model, prosody, syllable, articulation

1. Introduction

The Converter/Distributer model (C/D) model (Fujimura 2000) converts and distributes the phonological information to articulatory information. That is, it relates the abstract phonological, prosodic representation to the actual articulatory movements. Figure 1, adapted from Fujimura (1994), shows the overall organization of the C/D model. This figure is reproduced here to show that the C/D model is a comprehensive model of the phonology-phonetics mapping. This paper, however, focuses on a small but fundamental part of the model (shown in box): how to determine syllable prominence, timing and boundaries, or in short, "prosodic component" of the C/D model[2)]. Aspects of the C/D model that we set aside include how the C/D model's phonological representations look like, how consonantal and vocalic gestures are distributed to actual articulators (Distributor), and how they are actuated in actual articulation (Actuator). Instead, we focus on illustrating how syllable triangle representations (shown with a box in Figure 1) can be computed from actual articulation data.

Since the C/D model is a theory to map one representation (phonological) to another (articulatory), it allows us to both (i) predict articulatory patterns given some prosodic structures (from phonology to phonetics), and (ii) calculate prosodic structures given a set of actual articulatory data (from phonetics to phonology). The first component of the theory—the mapping from phonology to phonetics—has been tested in some of our previous research (Erickson et al. 2012, Erickson et al. 2014b); this paper illustrates the second feature of the model: calculating syllable triangle representations from actual articulatory data. Although this aspect of the C/D model is yet to be further explored, some preliminary results are reported in Erickson et al. (2015), Fujimura (2000), Kim et al. (2014) and Kim et al. (2015).

2. Constructing Syllable Triangle Representations

2.1 Syllable Magnitude

In the C/D model, phonological representations consist of a sequence of syllables, which are themselves made out of a nucleus (vowel) and onset and coda elements. An utterance is made up of, among other things, a train of syllable pulses that vary in height (=the

* Kanazawa Medical University（金沢医科大学）
** The Keio Institute of Cultural and Linguistic Studies（慶應義塾大学言語文化研究所）

syllables' magnitudes), as schematically illustrated in Figure 2.

The phonetic realization of the syllable magnitude is the degree of jaw displacement, which can be measured as a distance from the biteplane and the lower incisor[3]. Erickson et al. (2012) show that there is indeed a close correlation between the phonological prominence of a particular syllable and the degree of jaw displacement in English[4]. Kawahara et al. (2015) show that even in Japanese, some syllables within a phrase are prominent, and that they show large jaw displacement patterns.

2.2 Locating Syllable Triangles

In the C/D model, the magnitude of the syllable is determined by maximum jaw displacement. Locating the timing of the syllable pulse is slightly more complicated. The syllable is centered relative to the speed (maximum and minimum velocity) of the crucial articulators of the consonants. 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[5] velocity time function, the center of the syllable is defined as the midpoint between the syllable onset iceberg to the syllable coda iceberg (Bonaventura and Fujimura 2007, Fujimura 1986, 2000).

There are currently two proposed methods to identify iceberg points, each of which has its own virtues, and investigation is underway to compare these two methods[6]. To briefly describe each method, in the first type of method—which is actually what is first envisioned by Fujimura—the velocity measurements for these crucial articulators were made at the point where the velocity shows the least amount of change, i.e., where it is most stable. Fujimura (1986) observed that when one overlays the demisyllabic velocity time function, there is a point of smallest mean invariance, which he referred to as the iceberg region. The iceberg point is algorithmically determined at the minimum variance point of a number of trajectories of the same demisyllable (Bonaventura and Fujimura 2007, Fujimura 1986, 2000). We then find the point of the minimum root-mean-squared error in the horizontal direction after optimal time shifting of the trajectories to the reference trajectory (Bonaventura and Fujimura 2007, Fujimura 1986). Alternatively, one could choose the point of the minimum "iceberg metric" among multiple vertical movement bands of the crucial articulator (Menezes 2003). The iceberg metric is proportional to the variance of articulatory speed and inversely proportional to the mean of articulatory speed in the band. The method illustrated in this paragraph involves overlaying a large number of repetitions of the same sentence, including sentences with different emphasis

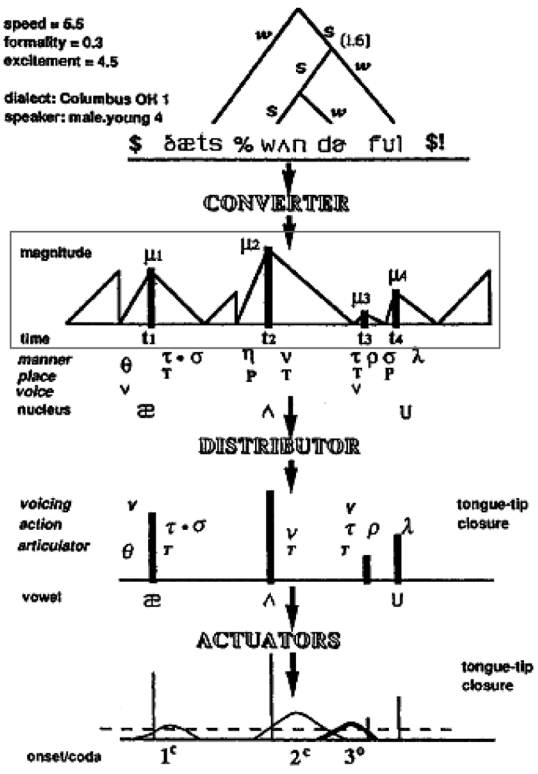

Figure 1 The C/D model: Overall organization (adapted from Fujimura 1994).

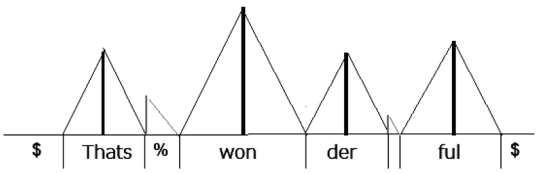

Figure 2 A sample syllable triangle representation (adapted from Fujimura and Erickson 2004).

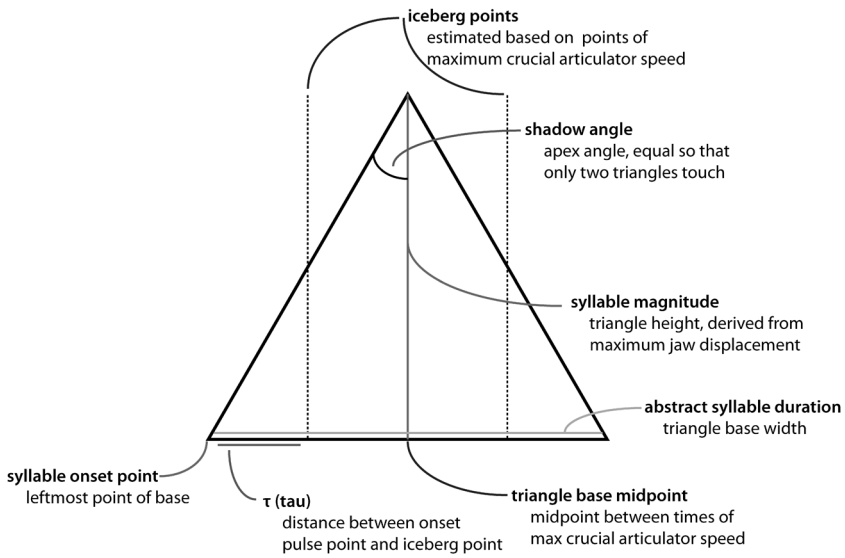

Figure 3 Illustration of how to create a syllable triangle (courtesy of Jeff Moore).

and phrasing patterns (e.g. Menezes 2003) in order to secure the reliability across different utterances; the goal of this method is to assess speaker characteristics across a large number of styles of speaking (P. C. Caroline Menzes 2015).

The second method to determine the point of least variance is to use the maximum speed point of the crucial articulators for the onset or coda of each demi-syllable, a method deployed by Erickson (2010) and Erickson et al. (2015). Although the method proposed by Fujimura is more accurate than taking the average of each curve's point of maximum speed (which is susceptible to large effects of small noise in the measurement of the actual data—see also note 6), it may be that both methods give similar determinations of articulatory syllable centers. Remember that one of the practical goals of locating iceberg (minimal variance) regions is to use these points as "anchor points" for determining the articulatory syllable center. From these points, then, syllable pulses (syllable triangle heights) are located, and this is what gives us the syllable timing for a string of syllables. As stated above, how comparable these two methods are is under investigation.

2.3 Construction of Syllable Triangles

To summarize the steps so far, the height of the syllable pulse is the amount of jaw displacement for each syllable, as measured from the biteplane. Placement of the syllable pulse is such that it occurs in the articulatory center of the syllable, determined as being halfway between the "iceberg" points of the syllable onset and offset crucial articulators. The next step is to construct actual syllable triangles.

Figure 3 illustrates the calculation of syllable triangles, which follows the following steps: (1) the apex of each triangle is the syllable magnitude, i.e., amount of jaw displacement; (2) the placement of the syllable pulse is the midpoint between the times of two iceberg points or, alternatively, of maximum crucial articulator velocity; (3) one constant angle, called "shadow" angle, is calculated 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. The syllable onset point is the left point of the base. The gap between two edges of adjacent triangles is the duration of prosodic boundary between the two. To implement these processes, one can use the automated algorithm in the UBEDIT software developed by Bryan Pardo (Menezes 2003, 2004, see also Erickson et al. 2015).

2.4 An Example

Let us now apply the computational steps described above to actual articulatory data. For the sake of simplicity, we deploy the simpler method to locate syllable triangles. Figure 4 shows articulatory data for the utterance *Pam said BAT that fat cat at that mat* (where *BAT* is contrastively emphasized) along with the constructed syllable triangles.

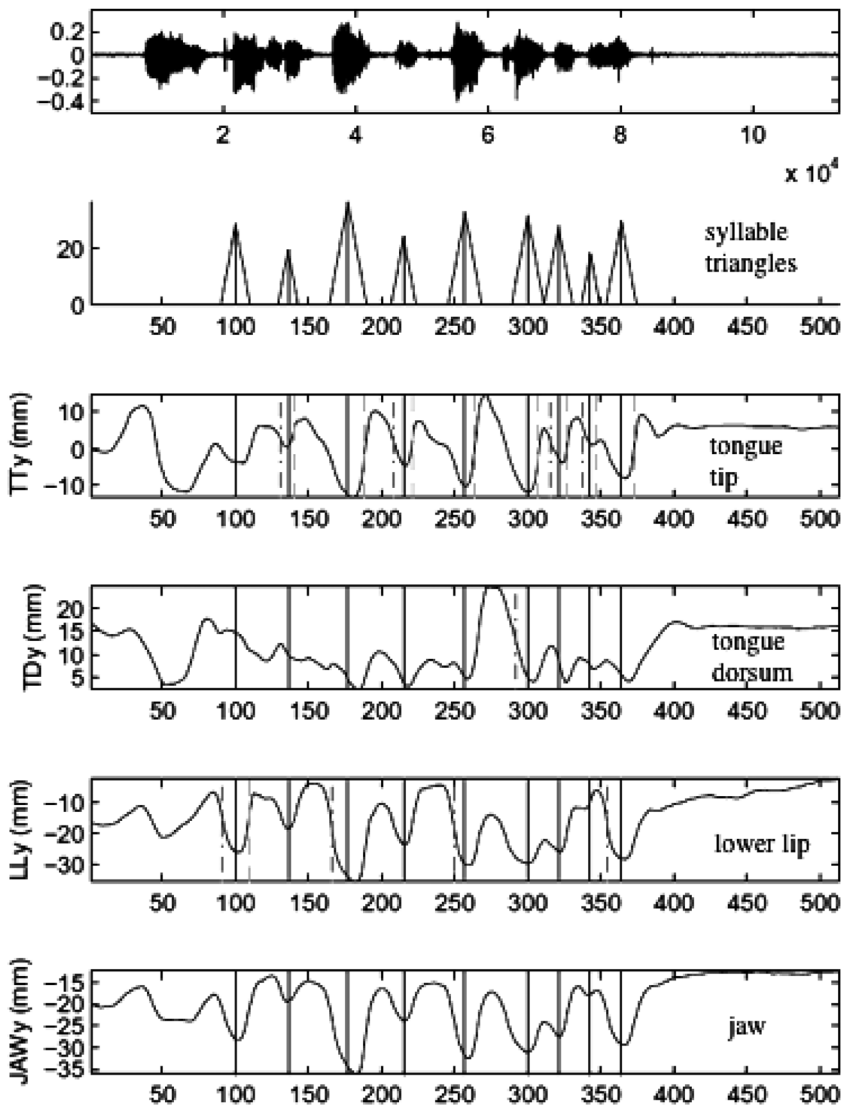

Figure 4 Syllable triangles constructed for the utterance *Pam said BAT that fat cat at that mat*, where contrastive emphasis is on *BAT* (Speaker A00, ut. 11), The top panel is the speech wave form, then the syllable triangles, then the tongue tip, then tongue dorsum, then lower lip, and bottom is jaw. In the panels showing the crucial articulators for the demisyllabic syllable onset and coda articulation, the dash-dot line denotes the iceberg time point for the onset; the dashed line denotes the iceberg time for the coda; the solid line in each of the bottom four panels denotes the articulatorily-based syllable center point.

In this figure, there are 9 vertical lines; each line represents the articulatory center of each of the 9 syllables in the utterance. The dashed lines indicate where the maximum speed occurs for the onset crucial articulator; the dash-dot line, that of the coda crucial articulator. The solid vertical line is for the word *Pam*: The crucial articulators for *Pam* are the lower lip. The solid line is placed half-way between the dash and dash-dot lines of the LLy (=lower lip). Notice that the solid line does NOT occur at the point of maximum jaw displacement (nor do any of the other solid lines which indicate the center of the syllable). The second solid vertical line is for the word *said:* the crucial articulators for *said* are the tongue tip, and the solid line is placed half-way

between the dashed and dot-dashed lines of the TTy (＝tongue tip). The third solid vertical line is for the word *bat:* The crucial articulators for *bat* are the lower lip for the onset, and the tongue tip for the coda, and the solid line is placed half-way between the dashed lines of the LLy and the dot-dashed lines of the TTy. The ensuing vertical solid lines are calculated in the same manner.

From these calculations, a syllable pulse train is constructed to represent the spoken utterance, as shown in the second panel of Figure 4. The gap between two edges of adjacent triangles (second panel, Figure 4) is the duration of the prosodic boundary between the two. In other words, the spaces between syllables show (a) where a boundary occurs and (b) how long the boundaries are.

This aspect of the C/D model—being able to calculate phrasal boundary durations from jaw movement data alone—makes the C/D model an empirically testable theory. The model makes specific predictions about how long phrasal boundaries should be given concrete articulatory data. Our preliminary experiment shows that this prediction is on the right track (Erickson et al. 2015), but further studies should be conducted.

Notes

1) This work was supported by NSF IIS-1116076, NIH DC007124, and the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (A)#22520412 and (C)#25370444 to the first author and JSPS grants #26770147 and #26284059 to the second author. A special acknowledgement to Jangwon Kim and Jeff Moore for their help with this paper.

2) In Fujimura and Erickson (2004), the prosodic component discussed here is referred to as "the time series of triangles [which] represents the skeleton of the base function of this utterance."

3) EMA is a useful tool to quantify the degrees of jaw displacement (see e.g. Erickson et al. 2012). In some cases, it is necessary to factor out the effect of vowel height to directly see the prosodic effects, because vowel height also affects jaw displacement in addition to prosodic strength (see e.g. Kawahara et al. 2014, Menezes and Erickson 2013, Williams et al. 2013). The C/D model's approach to prosody has focused on jaw displacement; F0 control of prosody is an aspect of the C/D model which is still "under construction." (p.c. Osamu Fujimura, 2015).

4) Along these lines, a great deal of prior research has also shown increased jaw displacement for emphasized, focused, contrastively emphasized syllables (e.g. de Jong 1995, Erickson 1998, 2002, Harrington et al. 2000, Macchi 1985, Menezes 2003, 2004, Stone 1981, Summers 1987, Westbury and Fujimura 1989, etc.). Previous studies have shown strong correlations between jaw displacement and F1 (e.g. Erickson et al. 2012, 2014a, 2014b, Kawahara et al. 2015, Kawahara et al. (submitted) and Menezes 2003, 2004). But note also they report no strong correlation with duration, intensity, and F0 in Japanese (Kawahara et al. submitted) or with F0 in English (Erickson et al. 2014b). The acoustic characteristics of increased syllable magnitude within the framework of the C/D model are a topic for a future paper.

5) Demi-syllables coincide with, in more traditional parlance, CV-sequences and VC-sequences given a CVC syllable. We use the term "demi-syllables," as the C/D model is crucially syllable-based.

6) After we finished this draft, Osamu Fujimura explained why he prefers the former method. To quote him, "My algorithm cannot determine the representative time value directly by observing moving speed as such if just that particular utterance is observed. My approach is based on the observation that a demisyllable movement of utterances (by a given speaker in a given utterance style) shares a relatively invariant movement pattern in a particular portion of the movement, which is determined as a specific part in terms of the active articulator's position in movement that shows convergence of the variance. Since it is based on an algorithm of setting a fixed critical threshold position that is crossed by individual demisyllabic movements for maximum convergence (in terms of instantaneous speed, that is the slope of the curve), the threshold position that provides the measure of crossing time for each movement pattern is inherently statistically defined relative to anatomical structure of the speech organs. Once this valid threshold position is set for a given set of utterances, the time values of demisyllabic occurrences for individual utterances are evaluated.

The reason for this algorithmic choice is two-fold. One is that it is more robust against the effect of small noise, which I expect in all physiological events. The other reason is, in general, evaluating time values of extrema, maximum, minimum, or inflection point as moving patterns is inherently difficult as a concept to capture. An extremum, by definition, is a point where any marked change is not found. It is hard to imagine that the natural biological system senses a point of maximum speed of motion and use it for evaluating speech organisation in conversation (This obvious difficulty may have been the reason that traditionally segementalism chose to avoid demisyllables). Setting a positional threshold line and observing the movement to determine the time of its crossing such a threshold line is inherently easier and more accurate and feasible to identify, if the threshold position refers to some hard object

like the hard palate or collision plane of lips even if it is not an apparent point of discontinuity."

References

Bonaventura, P. and O. Fujimura (2007) "Articulatory movements and phrase boundaries." In M.-J. Solé, P. S. Beddor, and M. Ohala (eds.) *Experimental approaches to phonology*, 209–227. Oxford: Oxford University Press.

de Jong, K. (1995) "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation." *Journal of the Acoustical Society of America* 97(1), 491–504.

Erickson, D. (1998) "Effects of contrastive emphasis on jaw opening." *Phonetica* 55(3), 147–169.

Erickson, D. (2002) "Articulation of extreme formant patterns for emphasized vowels." *Phonetica* 59(2-3), 134–149.

Erickson, D. (2010). "More about jaw, rhythm and metrical structure." *Acoustical Society of Japan Fall Meeting*, 103.

Erickson, D., A. Suemitsu, Y. Shibuya and M. Tiede (2012) "Metrical structure and production of English rhythm." *Phonetica* 69(3), 180–190.

Erickson, D., S. Kawahara, J. Moore C. Menezes, A. Suemitsu, J. Kim and Y. Shibuya (2014a) "Calculating articulatory syllable duration and phrase boundaries." *International Seminar Speech Production 2014 (Cologne, Germany, May 2014)*, 102–105.

Erickson, D., S. Kawahara, J. C. Williams, J. Moore, A. Suemitsu and Y. Shibuya (2014b) "Metrical structure and jaw displacement: An exploration." *Proceedings of Speech Prosody* 2014, 300–303.

Erickson, D., J. Kim, S. Kawahara, I. Wilson, C. Menezes, A. Suemitsu and J. Moore (2015) "Bridging articulation and perception: The C/D model and contrastive emphasis." *International Congress of Phonetic Sciences 2015*, #0527.

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

Fujimura, O. (1994) "Syllable timing computation in the C/D model." *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Yokohama, Japan, September 1994, 519–522.

Fujimura, O. (2000) "The C/D model and prosodic control of articulatory behavior." *Phonetica* 57(2–4), 128–138.

Fujimura, O. and D. Erickson (2004) "The C/D Model for prosodic representation of expressive speech in English." *Acoustical Society of Japan Fall Meeting, Okinawa*, 271–272.

Harrington, J., J. Fletcher and M. E. Beckman (2000) "Manner and place conflicts in the articulation of accent in Australian English." In M. Broe and J. Pierrehumbert (eds.) *Papers in Lab.Phonology V: Language acquisition and the lexicon*, 40–51. Cambridge: Cambridge University Press.

Kawahara, S., D. Erickson, J. Moore, A. Suemitsu and Y. Shibuya (2014) "Jaw displacement and metrical structure in Japanese: The effect of pitch accent, foot structure, and phrasal stress." *Journal of Phonetic Society of Japan* 18, 77–87.

Kawahara, S., D. Erickson and A. Suemitsu (2015) "Edge prominence and declination in Japanese jaw displacement patterns: A view from the C/D model." *Journal of Phonetic Society of Japan* 19(2), 33–43.

Kawahara, S., D. Erickson and A. Suemitsu (submitted) "A quantitative study of jaw opening: An EMA study of Japanese vowels."

Kim, J., D. Erickson, S. Lee and S. Narayanan (2014) "A study of invariant properties and variation patterns in the converter/distributor model for emotional speech." *Interspeech* 2014, 413–417.

Kim, J., D. Erickson and S. Lee (2015) "More about contrastive emphasis and the C/D model." *Journal of Phonetic Society of Japan* 19(2), 44–54.

Macchi, M. (1985) *Segmental and suprasegmental features and lip and jaw articulations*. *Doctoral dissertation New York University*. (unpublished)

Menezes, C. (2003) *Rhythmic pattern of American English: An articulatory & acoustic study*. *Doctoral dissertation, Department of Speech and Hearing Sciences, The Ohio State University*.

Menezes, C. (2004) "Changes in phrasing in semi-spontaneous emotional speech: Articulatory evidences." *Journal of the Phonetic Society of Japan* 8, 45–59.

Menezes, C. and D. Erickson (2013) "Intrinsic variations in jaw deviation in English vowels." *Proceedings of International Congress of Acoustics, POMA* 19, #060253.

Stone, M. (1981) "Evidence for a rhythm pattern in speech production: Observations of jaw movement." *Journal of Phonetics* 9(1), 109–120.

Summers, W. V. (1987) "Effects of stress and final consonant voicing on vowel production: Articulatory and acoustic analyses." *Journal of the Acoustical Society of America* 82(3), 847–863.

Westbury, J. and O. Fujimura (1989) "An articulatory characterization of contrastive emphasis." *Journal of the Acoustical Society of America* 85(S1), S98.

Williams, J. C., D. Erickson, Y. Ozaki, A. Suemitsu, N. Minematsu, and O. Fujimura (2013) "Neutralizing differences in jaw displacement for English vowels." *Proceedings of International Congress of Acoustics POMA* 19, #060268.

(Received May 21, 2015, Accepted Oct. 12, 2015)