Bridging articulation and perception: The C/D model and contrastive emphasis

Donna Erickson, Jangwon Kim, Shigeto Kawahara, Ian Wilson, Caroline Menezes, Atsuo Suemitsu, Jeff Moore

Kanazawa Medical University, Japan, University of Southern California, USA, Keio University, Japan, The University of Aizu, Japan, The University of Toledo, USA, Japan Advanced Institute of Science and Technology, Japan, Sophia University, Japan

EricksonDonna2000@gmail.com, jangwon@usc.edu, kawahara@icl.keio.ac.jp, wilson@u-aizu.ac.jp
Caroline.Menezes@utoledo.edu, sue@jaist.ac.jp, jeffmoore.personal@gmail.com

ABSTRACT

This paper compares prominence that listeners perceive with actual articulatory prominence. We calculated phrasal boundaries from articulatory patterns using an algorithm of the C/D model, and compared those calculated boundaries with perceived boundaries. The jaw displacements, measures of prominence, were measured using EMA; articulatory boundaries were derived from a C/D model algorithm. The data is a set of English sentences that vary in the placement of contrastive emphasis. Perception data were obtained from listeners who were asked to evaluate syllable prominence and syllable boundaries for these sentences. The results indicate that perception of syllable prominence shows strong correlations with articulatory prominence, showing that jaw displacement can be a strong perceptual cue for syllable prominence. Further, perception of syllable prominence is also correlated with algorithmically-calculated articulatory syllable boundaries. These results encourage us to explore the relation between articulation and perception of language prosody in terms of the C/D model framework.

Keywords: *articulation and perception, prosody, jaw displacement, RPT, C/D model*

1. INTRODUCTION

What are the articulatory correlates of prominence and phrasing, and how do listeners perceive them? To address this question, this paper compares perceptual prosody—the patterns of prominence and boundaries that listeners perceive—with articulatory prosody—actual articulatory patterns of utterance prominence and phrase boundaries. Perceptual prosody is assessed by the method of Rapid Prosodic Transcription (RPT) [1]. Articulatory prominence is based on articulatory EMA data, and articulatory boundaries were calculated using an algorithm offered by the C/D model [2]. We show that these articulatory measures correlate well with the obtained perception patterns.

Our theoretical framework is the C/D model, which offers an explicit algorithm to map phonological structures onto articulatory movements. The model is named "the C/D model", as it models how phonological distributions are "Converted" (C) and "Distributed" (D) across different articulatory gestures.

The prosodic structure of the utterance is phonetically determined, in part, by jaw displacement patterns: strong syllables have more extreme jaw displacements compared to weak syllables [3-6], and especially, contrastively emphasized syllables have increased jaw displacement [7-13].

The C/D model represents the patterns of jaw displacement in an utterance with a “syllable pulse train” [2]. Specifically, the height of the pulses represents jaw displacement measures, and the location of the pulses is computed using an algorithm described below. A taller pulse on the pulse train indicates greater jaw displacement (i.e. more stressed syllable or a syllable with larger magnitude [2])[i].

A working hypothesis is that each language has a pattern of jaw displacements which reflects the prosodic/metrical structure of that language. (See e.g., [6] for English, [15] for Japanese and [16] for Mandarin; also work in progress about Spanish and French.) One question that is addressed in this paper is whether jaw displacement correlates with perceived metrical prominence in English.

In addition, the C/D model offers an objective tool to algorithmically derive articulatory phrase boundaries in a spoken utterance, e.g., where and how big the phrase boundary is. This aspect of the C/D model was first hinted at in [2], and was further explored in recent studies by [14, 17]. This current paper continues along these lines.

The details of the algorithm proposed in [2] are as follows. First, the magnitude of the syllable is determined by the maximum jaw displacement of the syllable. Second, the consonantal gestures of the

syllable onset and coda (specifically, the movement velocity) determine the timing of the syllable pulse. Third, the syllable pulse is centered relative to the maximum speed (maximum or minimum velocity) of the crucial articulators of the onset and coda gestures (not the maximum jaw displacement). Crucial articulator (CA) refers to that articulator (tongue tip, tongue blade, tongue dorsum, lower lip) that articulates the consonants. For example, the CA for [t] is tongue tip, for [p] is lower lip, and for [k] is tongue dorsum. Fourth, 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" [18,19]. In this paper we use the maximum velocity speed points to represent the smallest variance of articulation, as described in [14].

To test this algorithm empirically, this paper calculated articulatory prominences and boundaries for a set of utterances which vary in emphasis conditions, as spoken by two North American speakers of English. The hypotheses are (i) jaw displacement will increase on the emphasized word, (ii) prosodic boundaries (location and size) will change as a function of the emphasis condition and (iii) these patterns of articulatory displacements and boundaries will match perceptual judgments of listeners, based on the RPT perceptual evaluations.

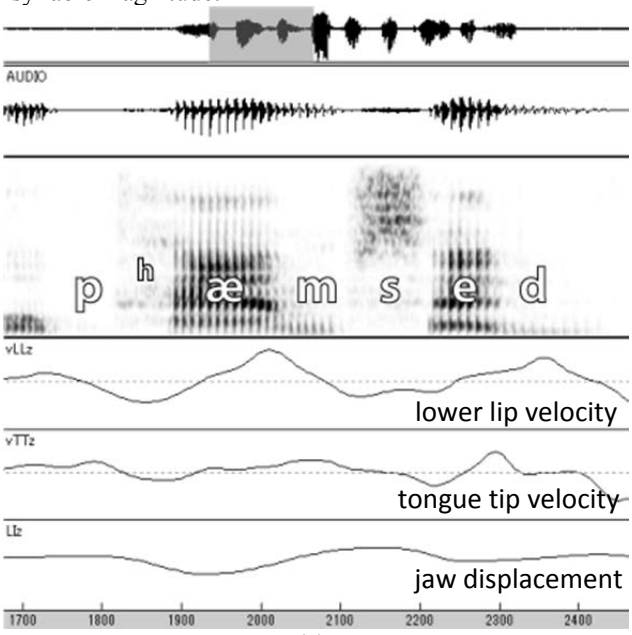
2. METHOD

The speakers were one female (A03) and one male (A05) North American English speakers. The utterances examined were (1) *Pam said bat that fat cat at that mat,* (2) *Pam said BAT that fat cat at that mat,* (3) *Pam said bat THAT fat cat at that mat,* (4) *Pam said bat that FAT cat at that mat,* (5) *Pam said bat that fat CAT at that mat*, where uppercase words indicate contrastive emphasis. Since jaw displacement varies as a function of vowel height, all syllables are closed syllables with [æ] vowels, or, in one case, [ɛ] (*said*). The utterances were presented to the speakers in randomized order, with five repetitions. The total number of utterances for A03 is 26, for A05 is 24.

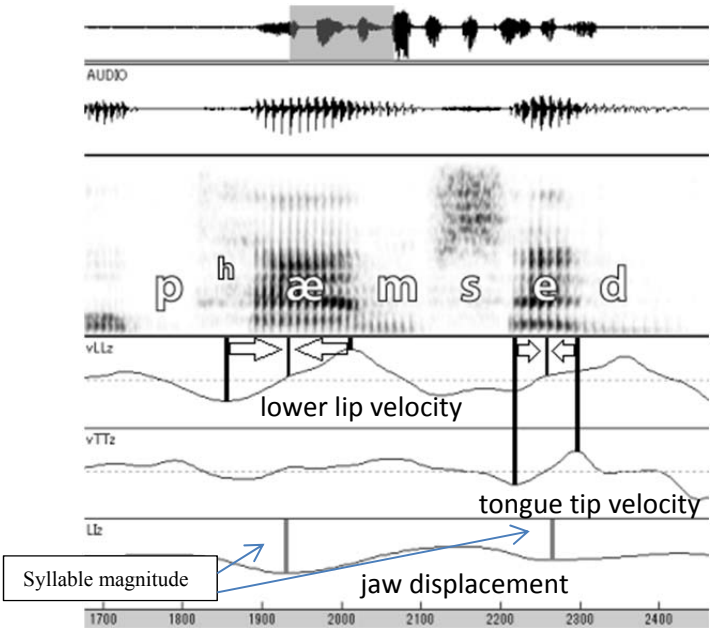
Acoustic and articulatory recordings were made using 3-D EMA (Carstens AG500). One sensor was placed on the lower medial incisors to track jaw motion, and one sensor each was placed on the tip of the tongue (TT), the mid of the tongue (TB) and the back of the tongue (TD). A sensor was placed on the lower 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 16 kHz, 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 of the jaw with respect to the bite plane was located for each target syllable of the utterance using the custom software mview (Mark Tiede, Haskins Laboratories). This measure was used to indicate the height of each syllable pulse (syllable magnitude) 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. The crucial articulators for the syllables in this sentence are lower lip [p, m, b, f], tongue tip [s, t, θ] and tongue dorsum [k].

Figure 1: (a) *Pam said* from an utterance (A03, ut 54) with no emphasis. Y-axis is jaw displacement (mm) or articulatory velocity (mm/s). The acoustic figures are shown on top, followed by Lower Lip (LL) Velocity, Tongue Tip (TT) Velocity, and Jaw Displacement. The peak velocity for LL for the [p] is 21.9 mm/s, for TT for the [s] is 13.1 mm/s, and the maximum jaw displacement for [ae] is 22.6 mm. In Figure (b) (next page) we mark vertical lines to show the points of maximum velocity for each of the Crucial Articulators for each of the words, and arrows pointing to the articulatory center of the syllable (marked with a thin vertical line). In addition, the amount of jaw displacement, in the bottom window, indicates the "syllable magnitude."



(a)

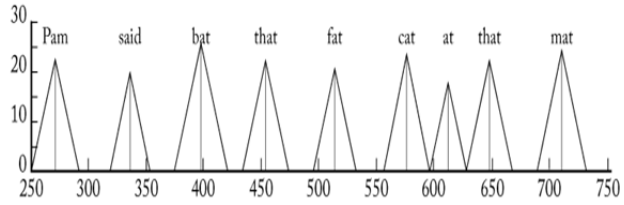


(b)

After the maximum jaw displacement and the articulatory center of the syllable were determined, we calculated the pulse train (pattern of syllable magnitudes) along with the articulatory boundaries. The MATLAB algorithm that implements this procedure is shown in the appendix.

The output is a series of syllable triangles, which represent the pulse train (syllable magnitude patterns) of the utterance. The height of each triangle is the amount of jaw displacement for that syllable; the center of each triangle is the articulatory center of the syllable, calculated as the midpoint between the maximum velocities of the onset and coda crucial articulators. The acute angle of the syllable triangle, namely, "shadow angle," for each syllable is computed, using the MATLAB script in the appendix, such that only two triangles (no more) touch. In this case, the point where triangles touch is between *cat* and *at.* (The distance between *at* and *that* is very small, but not zero.) The articulatory boundaries are defined as the distance between each triangle.

Figure 2: Syllable Triangles for Speaker A05, utterance 166, emphasis on BAT. Y-axis indicates jaw displacement in mm, x-axis is time in ms.



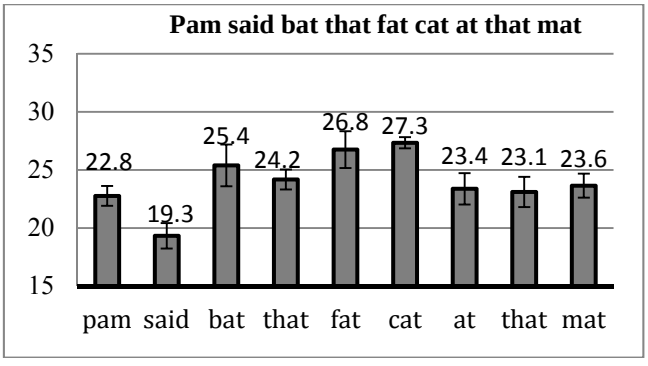
The articulatory parameters measured in this study are (1) the syllable magnitude (the height of each syllable triangle) and (2) the articulatory boundaries (the duration of the spaces between the triangles).

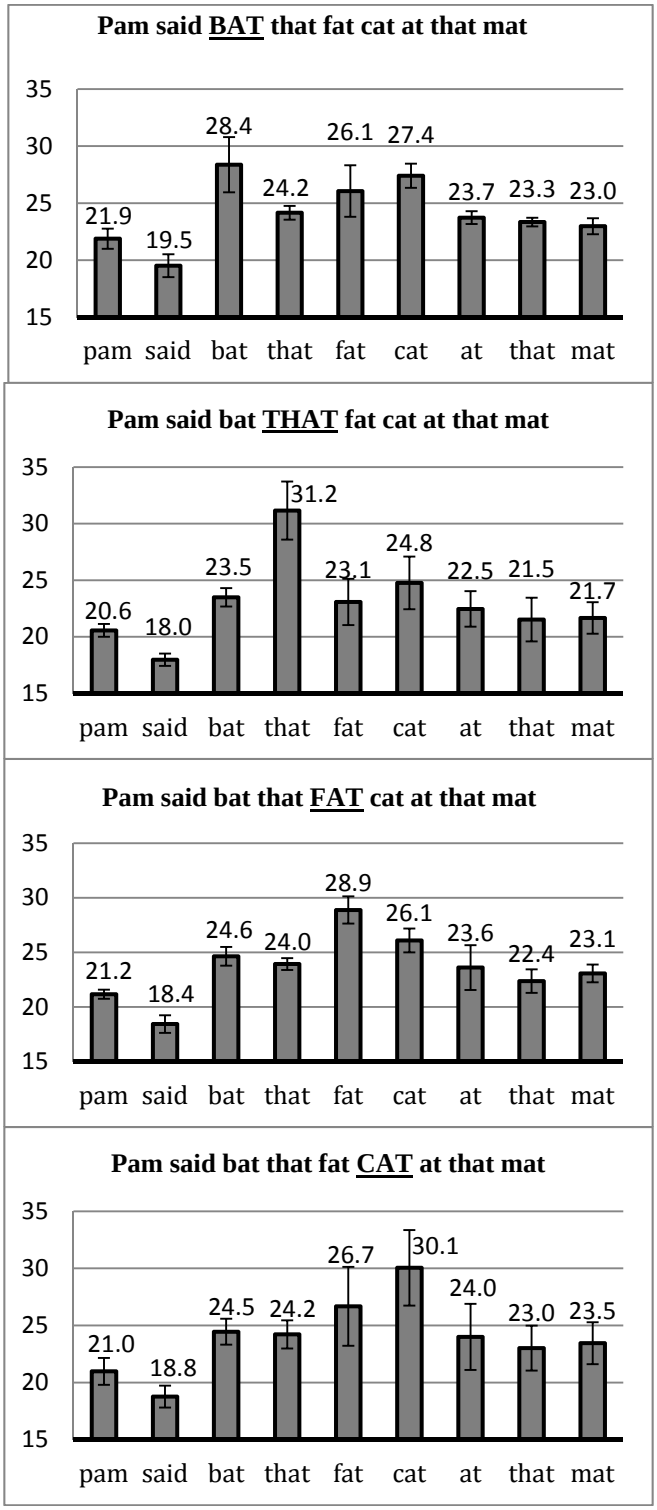
In order to assess if listeners can perceive the acoustic information generated by these articulatory patterns of syllable magnitudes and boundaries, an online Rapid Prosodic Transcription approach was used [20]. The participants were students at a midwestern city in the U.S. and a university in Tokyo. The students in Tokyo were advanced level English speakers, mostly Japanese who had lived a year or so in an English speaking country, plus one French and one Rwanda student. The instructions for the test were acoustically oriented [1]. L2 (Japanese) and L1 (English) listeners have been reported to perform similarly when doing RPT, given acoustic instructions [21].[ii] We have included the results of all listeners (N=18 for A03, and N=32 for A05). The instructions were to evaluate each sentence two times: the first time, they were to mark with a vertical line between a word where they heard breaks, discontinuities or pauses in the utterance; the second time, to mark with a line underneath the word if some words stand out more than others, if they are louder, longer or higher pitched. A sample sentence was given and listeners could listen to each sentence as often as they wished.

3. RESULTS AND DISCUSSION

The syllable magnitudes, as measured from jaw displacement for A03, are shown in Figure 3. A05 shows the same patterns but are not shown here due to space limitation. We observe that the emphasized word always has the largest syllable magnitude in an utterance, and also it is larger than the corresponding word in the non-emphasized version of the utterance. For example, the jaw displacement for bat in the neutral utterances is 25.4 mm and 28.4 mm in the emphasized one. Also, note the tendency for content words to have larger magnitude than function words. Sentences with emphasis on non-content words like "that", however, show a change in this pattern.

Figure 3: Averaged jaw displacement (mm, y-axis) for each syllable for A03.





The RPT results of the listeners' perceptions of syllable prominence and boundaries were tabulated, and Pearson correlation measures with articulatory measures were calculated. Table 1 shows the result of the correlation analysis, with Bonferroni-corrected *p*-values (0.05/4=0.0125). For both speakers, perceptual prominence and articulatory prominence, along with articulatorily-calculated boundaries, are well-correlated in general.

Table 1: Correlation analyses with perceived prosody vs. articulated prosody.

	Articulatory Prominence	Articulatory Boundaries
A03		
Perceptual Prominence	r=0.60 (p<.001)	r=0.36 (p<.001)
Perceptual Boundaries	r=0.43 (p<.001)	r=0.28 (p<.001)
A05		
Perceptual Prominence	r=0.68 (p<0.001)	r=0.41 (p<0.001)
Perceptual Boundaries	r=0.18 (p<0.05)	r=-0.18 *n.s.*

Articulatory prominence and perceptual prominence correlate very strongly for both speakers, suggesting that listeners' perception of prominence correlate well with jaw displacement magnitude. Articulatory boundaries and perceived boundaries show substantially weaker correlation; in fact, a non-significant reversal was found for Speaker A05. Articulatory boundaries, however, do show modest correlation with perceptual prominence.

This unexpected correlation can be explained as follows. There is a tendency for the articulatorily-calculated boundaries to increase in duration before the emphasized word: for Speaker A05, a modest correlation between syllable magnitude and articulatorily-calculated syllable boundary (r=0.33, p<.001) and a weaker one for A03 (r=0.21, p<0.025). In short, articulatory boundaries can be a good cue for articulatory prominence.

Then, it appears that listeners attend to the size of the articulatory boundary occurring before the word to aid in assessing if a word has prominence, in addition to attending to the size of the syllable magnitude.

4. CONCLUSION

Our findings show that (1) jaw displacement increases on the emphasized word, (2) articulatorily-calculated syllable boundaries tend to increase before an emphasized word, at least for one of the two speakers, (3) these patterns of jaw displacement/boundaries match perceptual judgments of listeners, based on the RPT perceptual evaluations. The present study also found supporting evidence for the hypothesis that articulatory prominence and algorithmically-calculated articulatory boundaries are significantly and positively correlated with perception of syllable prominence. It is interesting that articulatory boundaries give information to listeners about a sentence's prominence pattern.

5. ACKNOWLEDGEMENTS

This work was supported in part by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (C) #22520412 and (C) #25370444.

6. APPENDIX

Matlab script for calculating syllable triangle pulse train

```
function theta = f_comp_theta(syl_center_t, syl_mag)
num_syl = length(syl_center_t); % the number of syllables
tmp_theta_all = zeros(num_syl-1,1);
for which_theta = 1:(num_syl-1)
   tmp_theta_all(which_theta) = ...
      atan((syl_center_t(which_theta+1) -
syl_center_t(which_theta)) ...
      / (syl_mag(which_theta+1) + syl_mag(which_theta)));
end
theta = min(tmp_theta_all);
% plot syllable triangles
h=figure;
hold on;
for which_syl = 1:num_syl
   % draw a line for syllable magnitude bar
   line([syl_center_t(which_syl) syl_center_t(which_syl)],...
      [0 syl_mag(which_syl)]);
   % draw a line for the left line of syllable triangle
   line([syl_center_t(which_syl) (syl_center_t(which_syl) -
syl_mag(which_syl) * tan(theta))], ...
      [syl_mag(which_syl) 0]);
   % draw a line for the right line of syllable triangle
   line([syl_center_t(which_syl) (syl_center_t(which_syl) +
syl_mag(which_syl) * tan(theta))], ...
      [syl_mag(which_syl) 0]);
end
hold off;
```

7. REFERENCES

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

[2] Fujimura, O. 2000. The C/D model and prosodic control of articulatory behavior. *Phonetica* 57, 128-138.

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

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

[5] Menezes, C., Pardo, B., Erickson, D., and Fujimura, O. 2003. Changes in syllable magnitude and timing due to repeated corrections. *Sp. Communication* 40**,** 71-8**.**

[6] Erickson, D., A. Suemitsu, Y. Shibuya, and M. Tiede 2012. Metrical structure and production of English rhythm. *Phonetica* 69, 180–190.

[7] Macchi, M. 1985. Segmental and suprasegmental features and lip and jaw articulations. *Doct.diss. New York University (unpublished)*.

[8] Summers, W. V. Effects of stress and final consonant voicing on vowel production: articulatory and acoustic analyses. *J. Acoust. Soc. Am*. 82, 847–863.

[9] Westbury, J. Fujimura, O. 1989. An articulatory characterization of contrastive emphasis. *J. Acoust. Soc. Am*. 85, S98.

[10] Jong, K. de. 1995. The supraglottal articulation of prominence in English: linguistic stress as localized hyperarticulation. *J. Acoust. Soc. Am*. 97, 491–504.

[11] Harrington, J., Fletcher, J., Beckman, M.E. 2000. Manner and place conflicts in the articulation of accent in Australian English. In: Broe, M. Pierrehumbert, J. (eds), *Papers in Lab.Phonology V: Language Acquisition and the Lexicon*. Cambridge: Cambridge University Press, 40-51.

[12] Erickson, D., 1998. Effects of contrastive emphasis on jaw opening. *Phonetica* 55, 147-169.

[13] Erickson, D. 2002. Articulation of extreme formant patterns for emphasized vowels. *Phonetica* 59, 134-149.

[14] Kim, J., Erickson, D., Lee, S., Narayanan, S. 2014. A study of invariant properties and variation patterns in the converter/distributor model for emotional speech. *Interspeech 2014*. 413-417.

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

[16] Erickson, D., Iwata, R., Moore,J., Suemitsu, A., Shibuya, Y. 2015. The jaw keeps the beat: Speech rhythm in English, Japanese and Mandarin. *Lexicon Festa-3, Feb. 1, 2015. NINJAL*,Tokyo, Japan.

[17] Erickson, D., Kawahara, S., Moore, J., Menezes, C. Suemitsu, A., Kim, J., Shibuya, Y. 2014. Calculating articulatory syllable duration and phrase boundaries. *ISSP2014* (Cologne, Germany, May 2014), 102-105.

[18]Fujimura, O. 1986. Relative invariance of articulatory movements: An iceberg model. In: J. Perkell, J. and Klatt, D. H. (eds), *Invariance and Variability in Speech Processes*, Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. 226-242,

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

[20] http://gengojeff.netau.net/pam/

[21] Gabor P., Shinobu, M., Kazuhito Y. 2014. Boundary and Prominence Perception by Japanese Learners of English: A preliminary study. *Journal of Phonetic Society of Japan* 17, 59-66.

[i] Cold angry speech can show different jaw displacement patterns [14].

[ii] There may be some differences between L1 and L2 perception in RPT, but exploring these differences is a topic for future research.