A Question of Scope? Direct Comparison of Clear and In-Focus Speech Productions

Melissa A. Redford[1], Jessica N. Stine[1], Eric Vatikiotis-Bateson[2]

[1]*University of Oregon*
[2]*University of British Columbia*

redford@uoregon.edu, jstine@uoregon.edu, evb@mail.ubc.ca

Abstract

Parallels in the production of clear speech and words under prosodic focus suggest that both may be realized in the same way: as hyper-articulated speech. To directly investigate this possibility, school-aged children and college-aged adults produced target words in a default conversational style, a clear speech style, and with prosodic focus. The results were that children and adults both produced target vowels more distinctly and with greater mouth opening in the clear speech and prosodic focus conditions than in the default condition. Whereas the temporal scope of production changes varied as a function of condition in adults' speech, there was no evidence of this in children's speech.

Keywords: speech style, prosody, acquisition

1. Introduction

Clear speech has been characterized as hyper-articulated based on acoustic and perceptual findings of increased phonemic vowel contrast relative to a default conversational or casual speech style (Lindblom, 1990; Johnson et al., 1993). Emphatic stress due to prosodic focus has been similarly characterized based on acoustic and kinematic evidence that suggests more extreme vowel articulation for in-focus productions (de Jong, 1995). A clear speech style is also typically characterized by slower articulation rates (Picheny et al., 1986), just as prosodic focus is associated with lengthening (i.e., slowing; Turk & White, 1999). In spite of the parallels, clear speech and in-focus productions have yet to be directly compared. The current study makes this comparison to test the hypothesis that clear speech and in-focus productions are realized in the same way; the difference is only in the temporal scope of hyper-articulation, which is broader in clear speech. Because a single production strategy deployed for different linguistic ends would have advantages for acquisition, we chose to investigate the hypothesis in both child and adult speech.

School-aged children and college-aged adults produced target items embedded in meaningful sentences in a default conversational style, a clear speech style, and with prosodic focus. Vowel quality, movement of the lip-jaw complex, and durations associated with target word productions were analyzed as a function of speaking condition and age. The predictions were that phonemic vowel contrasts and peak displacements would be the same in the clear speech and in-focus conditions, and larger relative to the default condition. This prediction is consistent with the characterization of both clear speech and prosodic focus as hyper-articulated speech. Acoustic and kinematic durations of the word onset and vowel were expected to be longer relative to the total duration of the target word for in-focus and default productions than for clear speech productions, where an overall slowing effect was expected to lengthen word offsets. This prediction is consistent with the hypothesis that hyper-articulation has a broad temporal scope in clear speech. It is also consistent with the finding that the syllabic onset and nucleus are disproportionately affected by accentual lengthening in target words produced with prosodic focus (Turk & White, 1999). The effects of condition were expected to be the same in child and adult speech only if children control the settings that render hyper-articulated speech, and can appropriately vary the temporal scope over which these settings apply. This situation was deemed unlikely given evidence of prolonged speech motor control development (see, e.g., Walsh & Smith, 2002).

2. Methods

2.1. Participants

Sixteen native speakers of the standard west coast variety of American English participated in the study. Eight speakers were children (4 female), ranging in age from 7;2 to 7;8 with a mean age of 7;5. Children were in 2nd grade at the time of study and reading at grade level. The other 8 speakers (4 female) were college-aged adults.

2.2. Stimuli

The stimuli consisted of 16 meaningful sentences, each with a target word that occurred either early (subject position) or late (object position) in the sentence. All target words were adjectives that began with a bilabial stop, all modified a subsequent noun (*bag* or *boots*), and each contained an initial stressed syllable with 1 of 4 monophthongal American-English vowels (/i/, /æ/, /ɑ/, or /u/). Each vowel was represented by 2 target words: *beaded*/*peach*, *matte*/*black*, *modern*/*Prada*, *blue*/*puce*. Sentences were presented on cue cards. Font styling was varied as an additional cue to speech style. The cards were shuffled to create different random orders.

Table 1: *Example stimuli.*

Style	*Position*	*Example Sentence*
Default	Early	The beaded boots were brand new.
	Late	Mickey wanted the beaded boots.
Clear	Early	THE BLACK BOOTS WERE BARELY USED.
	Late	MATTHEW WANTED THE BLACK BOOTS.
In-focus	Early	The BLUE boots were barely used.
	Late	Billy wanted the BLUE boots.

2.3. Procedure

Lip-jaw movement was measured by tracking changes in mouth shape. To do this, 3 blue dots were placed at the vermillion border of the upper and lower lips and 1 at each corner of the mouth (top panel, Figure 1). Speakers were

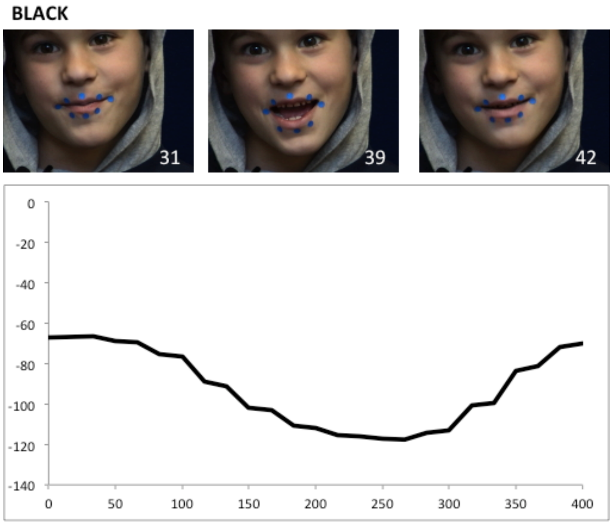

Figure 1: *Child participant producing target word "black" in the clear speech condition. Onset, offset, and peak opening frames are shown (top) along with the open-close movement in time (bottom). The y-axis in the bottom panel shows the distance between upper lip (UL) marker and lower lip (LL) marker, specifically, UL – LL; the x-axis, time in milliseconds.*

seated against a blue backdrop and in front of a video camera and music stand, which displayed the stimuli. The experimenter familiarized child speakers to the sentences by reading each of the cards out loud and with the child. All speakers quickly learned the sentences, which were similar in form (see, e.g., Table 1). After the initial familiarization phase, the experimenter stood to the side of the video camera. She interacted with the speaker throughout the experiment to help them render fluent sentence productions in the desired style.

Speaking style was blocked by condition. Sentences were first produced in a default conversational style, then in a clear speech style, and finally with the target adjective in focus. Clear speech was elicited by having the experimenter feign hard-of-hearing and asking repeatedly, "*What did you say*?" In-focus renditions were elicited by asking the speaker to correct the experimenter's 'inaccurate' reading of a sentence, which was produced with questioning intonation; for example, "*Billy wanted the peach boots*?" to which the speaker would appropriately reply, "*No, Billy wanted the* <u>*BLUE*</u> *boots*." Quick, quiet oral readings of each sentence were given during the task itself if a child needed help to fluently read/speak a sentence. Sentences were repeated if a first rendition was halting or non-fluent. Audio-video recording using a SONY DCR-PC101 camera captured lip-jaw movement. Higher-quality audio was recorded separately to a Marantz PMD660 using a Shure ULXS4 standard wireless receiver and a lavaliere microphone, located in a fixed position a few inches from the speaker's mouth.

2.4. Measurement

Acoustic measurements were made in Praat (Boersma & Weenink, 2011). Utterances were displayed as an oscillogram and spectrogram. Target word durations were extracted based on audio-visual inspection of the waveform for cues to the stop closures that delimited word onsets and offsets. The duration of the initial C(C)V sequence was also taken, with vowel offsets easily identified by amplitude and spectral changes associated with closure for the post-vocalic consonant, all of which were obstruent consonants. Formant measures were extracted at the midpoint of F2 during its steadiest portion. Once acquired, formant values were converted from Hertz to Bark (z) using the formula proposed in Traunmüller (1990). Formant values were then normalized across speakers using a Bark difference metric (e.g., Syrdal and Gopal, 1986); specifically, z3 – z1 for information regarding degree of vocal tract stricture, and z3 – z2 for information regarding tongue advancement.

Mouth shape was extracted by exporting acquired color video as an image sequence at 30 frames per second. Barbosa and Vatikiotis-Bateson's (2006) algorithm was then used to track the position of the blue dots (i.e., markers) placed around the speaker's mouth. Position was defined according to the vertical and horizontal dimensions of the frame, measured in pixels. Movement was defined by the change in position of the markers from frame to frame. The analyses presented here focus only on opening during target word production, and thus on the difference between the medial upper and lower lip markers in the vertical (y) dimension of the 2D space (bottom panel, Figure 1). The frames associated with target word production were identified with reference to word-onset bilabial articulations: target word onset was defined as the first frame in which full bilabial closure was achieved, and target word offset as the frame preceding the next event of full bilabial closure (top panel, Figure 1).

2.5. Analyses

Linear mixed effects modeling was used to determine the effect of speaking condition (default, clear, in-focus) and age (child vs. adult) on the measures we use to assess hyper-articulation and its scope during target word production. Target word position (early vs. late) and vowel (/i/, /æ/, /ɑ/, or /u/), when relevant, were included as additional fixed effects in the analyses. Speaker and word, when relevant, were treated as random effects. Measures of hyper-articulation were vowel space size and maximum opening. Vowel space size was calculated as the perimeter of the quadrilateral defined by the 4 target vowels in the normalized formant space. Measures of scope were the relative duration of the initial C(C)V sequence to overall target word duration and the relative duration of the opening movement (time to maximum opening) expressed as a proportion of the open-close cycle (or total movement, if more than one cycle) associated with target word production. All results are reported with estimated denominator degrees of freedom rounded to the nearest whole number.

3. Results

3.1. Hyper-articulation

The first set of analyses investigated the effects of condition and age group on vowel contrastiveness to test the hypothesis that clear and in-focus productions are both realized as hyper-articulated speech. Consistent with this hypothesis, the analysis of perimeter values indicated that the vowel space was larger for clear and in-focus productions of the target words than for default productions [$F(2,70) = 8.97$, $p < .001$], as shown in Figure 2. The analysis also indicated a significant effect of position within the phrase [$F(1,70) = 5.87$, $p = .018$] such that the vowel space perimeter was slightly larger when target words modified the object noun than when they modified the subject noun. There was no significant effect of age group on the perimeter values calculated in normalized F1 × F2 space. In addition, there were no significant interactions between any of the fixed effects, indicating a consistent effect of condition regardless of the speakers' age or target word position within the sentence.

To investigate the extent to which individual vowels were articulated differently as a function of condition, analyses were also conducted on the normed F1 and F2 values associated with each vowel. These analyses indicated a significant effect of condition on the normalized F1 values for /i/ [*F*(2, 148) = 3.62, *p* = .029], and on the normalized F2 values for /ɑ/ [*F*(2, 144) = 7.27, *p* = .001] and /u/ [*F*(2, 149) = 11.87, *p* < .001]. There was also a significant effect of age on the normalized F2 values for /i/ [*F*(1, 14) = 15.27, *p* = .002]: children produced /i/ with higher F2 values (bark distance from F3 was smaller) than adults. Post hoc tests revealed no significant differences between clear and in-focus productions of /i/, /ɑ/, or /u/.

The analyses on maximum opening produced similar results to those on vowel quality. Consistent with the hypothesis of hyper-articulation, clear and in-focus productions of the target words resulted in greater maximum opening than default productions [*F*(2, 529) = 103.26, *p* < .001]. Not surprisingly, maximum opening also varied systematically with vowel [*F*(3, 101) = 131.36, *p* < .001]. The interaction between condition and vowel was also significant [*F*(6, 530) = 5.05, *p* < .001]. Analyses within each vowel nonetheless indicated that production varied systematically with condition regardless of the vowel in the target word [/i/, *F*(2, 140) = 37.28, *p* < .001; /æ/, *F*(2, 130) = 37.50, *p* < .001; /ɑ/, *F*(2, 126) = 20.96, *p* < .001; /u/, *F*(2, 139) = 14.80, *p* < .001]. Similarly, post hoc comparisons indicated larger opening values for clear and in-focus productions than for default productions. This was true for all target words except those with the high back vowel, where only clear speech productions were associated with significantly more opening than default speech productions.

The condition by age interaction on maximum mouth opening was also significant [*F*(2, 529) = 6.98, *p* = .001], even though the simple effect of age was not. When the analysis was split by speakers' age, production was still found to vary systematically with condition [child, F(2, 313) = 16.16, p < .001; adult, F(2, 320) = 91.80, p < .001]. Inspection of mean differences suggest that the interaction was due to the finding that children produced target words in the clear condition with somewhat greater opening values than those in the in-focus condition and vice versa for the adults (Figure 3). Variance in child productions was such, however, that the difference between clear and in-focus productions was only significant for the adults [mean difference = 2.57, *p* = .048].

3.2. Scope

The next set of analyses investigated the effects of condition and age group on the relative time devoted to articulation of the target word onset + stressed vowel sequence. The goal was to address the question of scope differences in clear and in-focus productions. The results were as follows.

The relative acoustic duration of the initial onset+vowel sequence in the target word varied systematically by condition [*F*(2, 607) = 28.45, *p* < .001] and, of course, by vowel [*F*(3, 120) = 50.14, *p* < .001]. The condition by vowel interaction was also significant [*F*(6, 607) = 3.39, *p* = .003], but within-vowel analyses nonetheless indicated that production varied systematically in spite of the interaction [/i/, *F*(2, 150) = 9.88, *p* < .001; /æ/, *F*(2, 155) = 6.27, *p* = .002; /ɑ/, *F*(2, 150) = 14.70, *p* < .001; /u/, *F*(2, 152) = 5.35, *p* = .006].

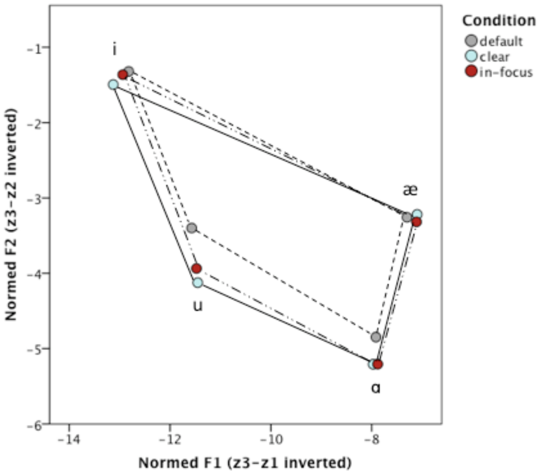

Figure 3: *Mean normalized formant values for the 4 monophthongal vowel targets are shown as a function of speaking condition. The lines that connect the vowels define the perimeter of the vowel space, providing a measure of phonemic contrast.*

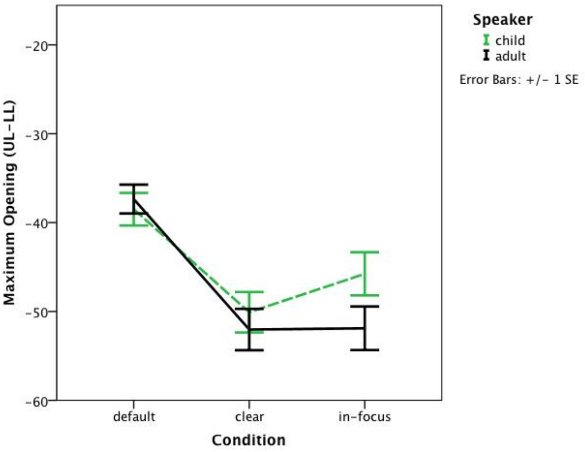

Figure 3: *Maximum opening values are shown as a function of condition and age group. Larger negative values indicate greater vertical distances between the upper and lower lip markers.*

Although the effect of age was not significant in the overall analysis, the interaction between condition and age was [*F*(2, 607) = 13.50, *p* < .003], as shown in Figure 4. Post hoc mean comparisons confirmed the differences evident in the figure: children produced longer onset+vowel sequences when the target words were under prosodic focus (mean difference from clear = -.02, *p* = .017; mean difference from default = -.03, *p* < .001) than when they were spoken in a clear or default speech style, but children did not differentiate the time to articulate onset+vowel sequences in the clear and default speech styles. Adults also produced the longest onset+vowel sequences in target words under prosodic focus (mean difference from clear = -.05, *p* < .001; mean difference from default = -.02, p = .004), but the clear speech onset+vowel sequences were shorter relative to the duration of the whole word than the default speech onset+vowel sequences (mean difference from default = .03, *p* < .001).

The differences between child and adult productions were even more striking in the analyses of the relative opening duration. Again, there were significant effects of condition [*F*(2, 539) = 4.05, *p* = .018] and vowel [*F*(3, 107) = 4.09, p = .009] on the duration of the opening phase, but no significant interaction between them. There was also a significant effect of age [*F*(1, 14) = 22.71, *p* < .001]: the opening and closing phases in adults' productions of the target words were more

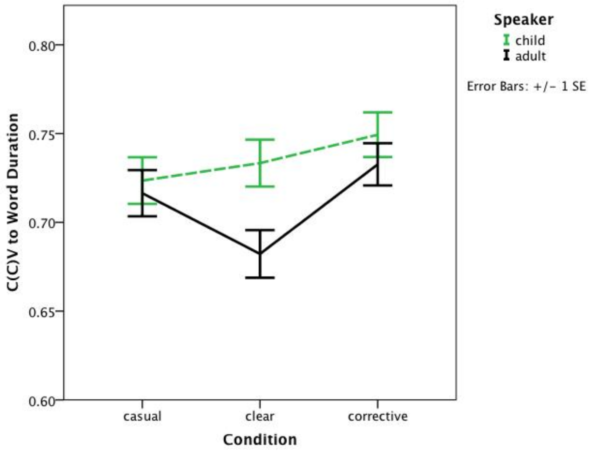

Figure 4: *The relative acoustic duration of onset sequences in the target word is shown as a function of speaking condition and age.*

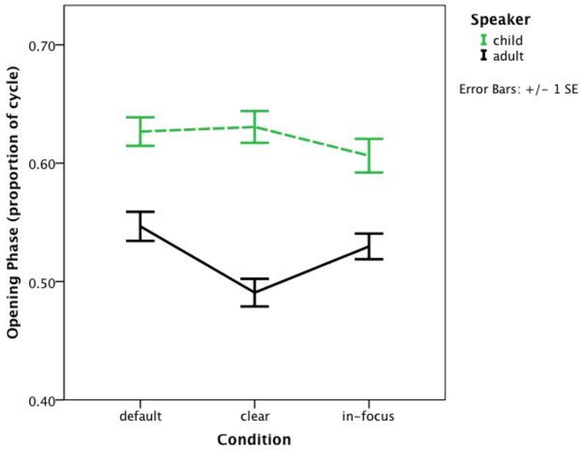

Figure 5: *The relative time to peak opening is shown as a function of speaking condition and age.*

symmetric than in children's productions. With respect to the question of scope, the overall analysis indicated a significant interaction between condition and age [*F*(2, 538) = 6.08, *p* = .002], which is shown in Figure 5. Analyses on data split by speakers' age confirm the pattern evident in the figure; namely, the effect of condition was significant only in adults' speech [*F*(2, 271) = 10.84, *p* < .001]. As with acoustic durations, adults' opening phase was relatively shorter in clear speech productions of the target word compared to in-focus or default speech productions (mean difference from in-focus = .04, *p* = .001; mean difference from default = .06, *p* < .001). Also, the difference between in-focus and default speech productions was not significant in the adult data.

4. Discussion and conclusion

The results on vowel quality and movement of the lip-jaw complex are consistent with the hypothesis that both clear speech and emphatic stress due to prosodic focus are realized as hyper-articulated speech. The vowel space was equally large in the clear speech and in-focus conditions. Analyses on normalized F1 and F2 also revealed no significant differences between clear and in-focus productions of /i/, /ɑ/, or /u/ (/æ/ was not effected by speaking condition). Also, peak displacement was greater in the clear speech and in-focus conditions compared to the default condition. None of these results varied with the speakers' age, suggesting that 7-year-old children have acquired adult-like control over the parameter settings relevant for hyper-articulation.

The results on time spent in onset+vowel articulation were consistent with the hypothesis that it is the temporal scope of supraglottal articulatory changes that distinguishes clear speech from in-focus production. As expected, onset+vowel articulation was shorter relative to whole word articulation for clear speech productions compared to in-focus or default productions, consistent with a broader scope of change in clear speech relative to in-focus speech. However, unlike the results on hyper-articulation, the temporal results varied significantly with speakers' age. Children spent more time overall on the onset+vowel sequence than adults, probably because they were less adept at the articulation of the complex onsets in the target word stimuli. More intriguingly, children's emphasis on opening was observed regardless of speaking condition. This result could indicate less fine-grained temporal control over different modes of production in children. More generally, it suggests that the development of timing control is more protracted than control over parameters such as stiffness that underlie articulatory changes associated with hyper-articulation.

5. Acknowledgements

This research was supported by award #R01HD061458 from NICHD. The authors thank Faire Holliday for her significant help with data processing.

6. References

Barbosa, A. V., and Vatikiotis-Bateson, E. (2006). Video tracking of 2D face motion during speech. In: *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology – ISSPIT' 2006, Vancouver, Canada*, pp. 791-796.

Boersma, P., and Weenink, D. (2011). Praat: doing phonetics by computer (Version 5.2.11). Retrieved January, 18, 2011.

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

Johnson, K., Flemming, E., and Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language*, 505-528.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In: *Speech production and speech modelling* (pp. 403-439). Springer Netherlands.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1986). Speaking clearly for the hard of hearing II. *Journal of Speech, Language and Hearing Research*, 29, 434.

Syrdal, A.K., and Gopal, H.S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, 79, 1086–1100.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America*, 88, 97–100.

Turk, A. E., and White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics*, 27, 171-206.

Walsh, B., and Smith, A. (2002). Articulatory movements in adolescents: Evidence for protracted development of speech motor control processes. *Journal of Speech, Language, and Hearing Research*, 45, 1119-1133.