ARTICULATORY MEASURES OF PLANNED AND UNPLANNED COARTICULATION

Argyro Katsika[1], D. H. Whalen[1,2,3], Mark K. Tiede[1], Hannah M. King[1],

[1]Haskins Laboratories, [2]Graduate Center of the City University of New York, [3]Yale University
argyro.katsika@yale.edu, whalen@haskins.yale.edu, tiede@haskins.yale.edu, hannah.king@yale.edu

ABSTRACT

The extent and sources of coarticulation remain an active topic of research. Whalen [18] found that vowel-to-vowel coarticulation occurred even when the utterance was initiated before an intervening consonant's identity was known. Here, we address vowel-to-vowel coarticulation in the articulatory domain, replicating [18]. Spoken English nonsense strings [əˈbVCɑ] were recorded with electromagnetic articulography. The V was [i] or [u] while the C was [b] or [p]. In one condition, the V was known but the C was not, it was reversed in the other condition. The missing information appeared once phonation began. Our results revealed speaker-dependent motor control strategies, with anticipatory effects of the V on the tongue position of the initial schwa being present for speakers who anticipated the V identity, but not otherwise. These patterns are consistent with the position that coarticulation is planned, corroborating Whalen's [18] conclusions.

Keywords: coarticulation; articulation; planning; motor control.

1. INTRODUCTION

Segments in a linguistic utterance influence each other in systematic but complicated ways through the process of coarticulation [9, 11]. This overlap is useful to human listeners because it allows them to convey sounds more quickly than would be possible without such overlap. While this is helpful to humans, it has proven to be a major obstacle to automatic processing. [7, 13]. Automatic speech recognition (ASR) has improved greatly in recent years, but it still has difficulty handling changes in speaker or dialect [14] or with speech produced in noisy environments [16]. If we can better understand the operation of coarticulation, we can incorporate that knowledge into ASR systems.

Listeners not only interpret the coarticulation in speech, they depend on it [4, 6, 15]. The effects of vowel place and lip rounding on fricative noises also provide information for vowels, and the acoustic feature of duration provides information about vowel quality and consonant voicing category at the same time [17].

Coarticulation was once described as the formant transitions immediately following the release of a consonant constriction [10] but is now more widely understood as the influence of one segment on another. While adjacent segments clearly influence one another, segments that are separated by one or more additional segments can be affected as well [2, 3]. The formants of vowels can come to resemble those of another vowel even if they are separated by a consonant [1, 12]. Whalen [18] examined the acoustic signal for evidence of the relative influence of planning and inertia. The current study re-examines this issue using articulatory measures.

The methodology of [18] involved presenting a nonword of the form [eˈbVCɑ] to a speaker on a computer screen, but in every stimulus one letter was missing. Sometimes the missing letter was the V, and sometimes it was the C. The speaker prepared the utterance and then began speaking. Once phonation began, the letter representing the missing segment (either a consonant or a vowel) appeared on the screen, and the speaker attempted to complete the utterance as naturally as possible. Coarticulation was examined both before the C or V (anticipatory) or after it (perseverative). Anticipatory effects of the vowel were reduced or eliminated when it was unknown at initiation, but they were typical when it was known, even if the effect crossed a segment (the consonant) whose identity was not known. Perseverative effects were similar in all conditions. The results were interpreted as showing that planning was responsible for the large temporal extent of the anticipatory effects.

The present experiment has investigated the extent of planning involved in vowel-to-vowel coarticulation, using the same stimuli as Experiment 2 of [18], in which English nonwords of the form [əˈbVCɑ] were used. The V was [i] or [u] while the C was [b] or [p]. In one condition, the V was known but the C was not, with the reverse in the other condition. We predict anticipatory articulatory effects will be detected on the first vowel of the sequence when the second vowel is expected.

2. EXPERIMENT

2.1.1. Apparatus

Kinematic data were acquired using electromagnetic articulometry (EMA; WAVE, NDI). Speech movements were digitized at 100 Hz, and concurrently recorded audio was sampled at 44.1 kHz. Sensors were glued to the lips (upper, lower and mouth corner), tongue (tip, center and dorsum), upper and lower incisors, and left and right mastoids. The data were corrected for head movement and translated to the occlusal plane.

2.1.2. Participants

Three native speakers of English served as participants. Two were female (F01 and F02) and one, male (M03). They provided informed consent and were paid for their participation.

2.1.3. Stimuli and procedure

The experiment uses the same stimuli and follows the same general procedure as that of Experiment 2 of [18]. English nonsense strings [əˈbVCɑ] were used. The stressed V was [i] or [u], while the C was [b] or [p]. Either the V or the C was unknown at the beginning of each trial, yielding eight stimuli in total: four words ([əˈbubɑ], [əˈbibɑ], [əˈbupɑ], and [əˈbipɑ] * 2 levels of the Unknown factor (levels: unknown consonant, unknown vowel).

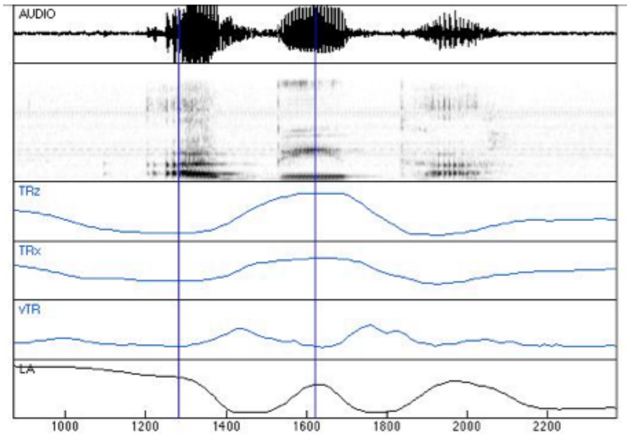
A training session familiarizing the participants with the task preceded the experimental session by 1-3 days. During the experimental session, for speaker F01, twenty blocks of the test stimuli were presented, each containing one repetition of the stimuli in a randomized order. For speakers F02 and F03, a longer training session was used and the number of experimental blocks was reduced to 12 to minimize the total length of the experiment.

Custom software (*Marta*, Haskins Laboratories) was used to control stimulus presentation, phonatory monitoring, and audio recording. The speaker was seated before a computer screen on which the stimuli were presented. For each trial, the stimulus was shown missing one letter (e.g., "UHBI_A"). The missing letter appeared once phonation exceeded a particular target amplitude threshold. The task of the speaker was to incorporate the new segment as smoothly as possible into the utterance, and to then repeat the word fluently with all information available. Here, the analysis of the words that involved segment incorporation is presented. Only correct productions of the test words were included in the analysis.

2.1.4. Measurements

The kinematic data were labelled using velocity extrema with *Mview* (Haskins Laboratories). Specifically, the point of maximal opening tracked by the tongue dorsum (TD) sensor during the production of the schwa and the V (i/u) were detected (Figure 1), since these were taken to be representative approximations of the articulatory targets of their respective vowels. For each of these points, the following pair of values was extracted: horizontal (from posterior to anterior) and vertical (from inferior to superior) TD displacement.

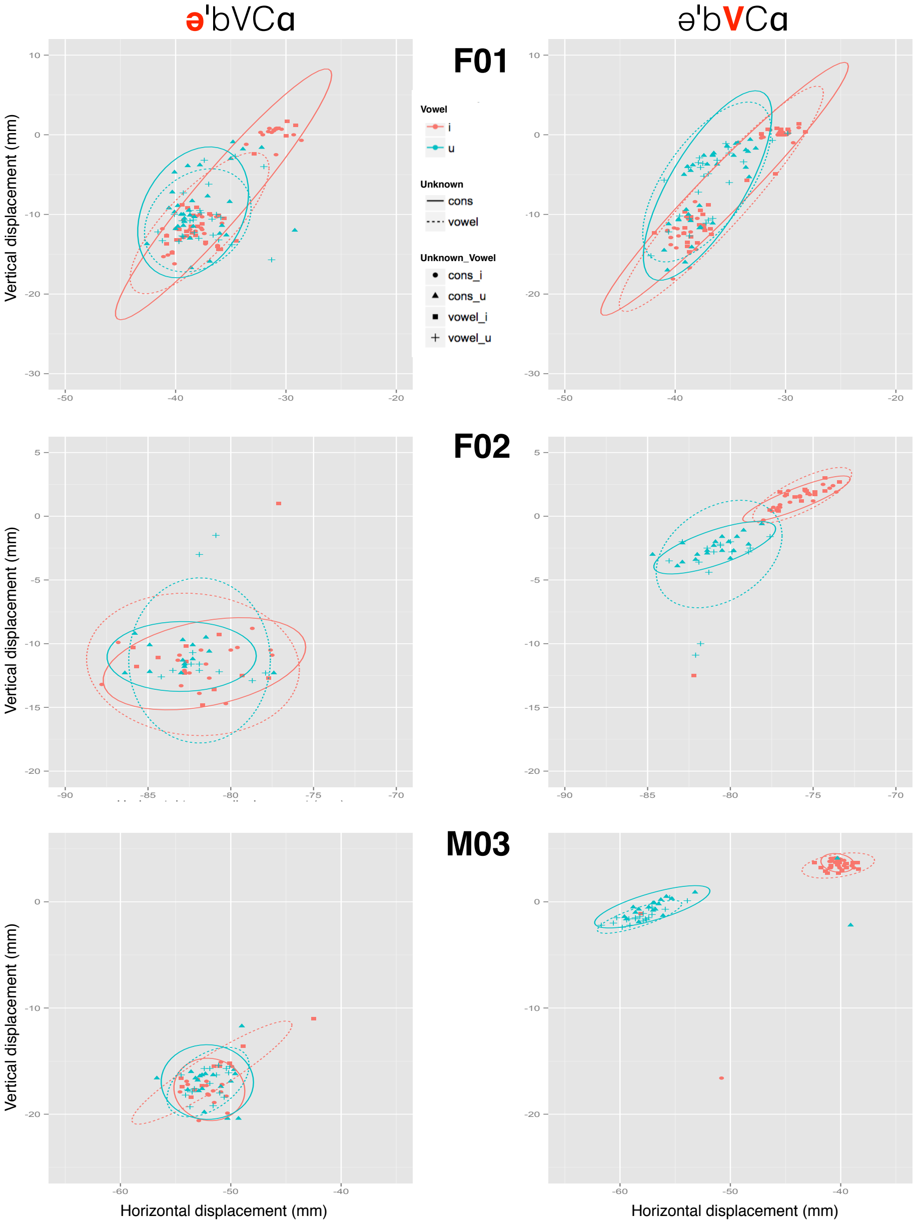
Figure 1: An instance of [əˈbiba], with the second C unknown, as uttered by speaker F01. The first two panels correspond to the audio signal (waveform and spectrogram). The third, fourth and fifth panels represent vertical displacement, horizontal displacement and velocity of the tongue dorsum respectively. The last panel represents lip aperture. The vertical lines across all panels show the points of maximum opening for the schwa ([ə]) and the V ([i]) respectively.



3. RESULTS

Figure 2 presents three pairs of plots, corresponding to speakers F01, F02 and M03 respectively. The left column shows the target positions of the tongue dorsum for the tokens of the schwa, i.e., the first vowel of the word. The right column displays the respective positions for the V, i.e., the second vowel, which is either [i] or [u] depending on the condition. Red represents [i], and blue [u]. Circles mean that the consonant is unknown and the vowel is [i] (cons_i), squares that the vowel is unknown and will be [i] (vowel_i), triangles that the consonant is unknown and the vowel is [u] (cons_u), and crosses that the vowel is unknown and will be [u] (vowel_u). The figure also includes 95% confidence ellipses superimposed on the four groups. Solid ellipses include tokens with the consonant unknown, and broken ellipses tokens with the vowel unknown.

Figure 2: Tongue dorsum position (in mm) (along with confidence ellipses) in the two-dimensional space at the articulatory target for the schwa (left column) and the V (right column) for each speaker separately. Horizontal displacement is presented from posterior (right of the x axis) to anterior (left of the x axis), and vertical displacement from inferior (bottom of the y axis) to superior (top of the y axis).



As Figure 2 reveals, speaker F01 shows a distinct pattern from speakers F02 and M03. Specifically, the quality of the following vowel (V) affects the TD position during the articulatory target for the schwa. The tongue dorsum is more anterior, and possibly more superior as well, when the following V is [i] than when the V is [u], especially in the consonant unknown condition (i.e., when the V is known). For the other two speakers (F02 and M03) on the other hand, the quality of the second vowel does not exert any effect on the TD position during schwa's articulatory target, regardless of whether the V identity is known (i.e., when the consonant is unknown) or not. These patterns are reversed when turning to the spatial profile of the V. For speaker F01, the V observations are not clearly classified into two distinct V qualities ([i] and [u]), although their identity was already known to the speaker by the time they were produced and they were heard as correct. Contrary to F01, speakers F02 and M03 have two distinct TD position categories corresponding to the front vowel [i] and back vowel [u] respectively.

The classification patterns shown in Figure 2 are further clarified by the means of two sets of linear discriminant analysis; one set assessing the V quality (i.e., whether V is either [i] or [u]) as a classifier of the TD position in the two-dimensional space during the schwa, and the other set during the V. Each speaker is examined separately. Tables 1 and 2 present the cross-validation results (shown as percentages) of these two sets of analyses respectively. Columns represent members of actual classes, and rows members of predicted classes. Correctly predicted instances are located in the diagonal cells. The quality of V is a fairly good predictor of TD position during the schwa for speaker F01 (59% success when the V is [i], and 62% when the V is [u]). For the other two speakers, the model performs better for one vowel context (72% when the V is [i] for F01, and 92% success when V is [u] for M03), but fails in the other (27% when the V is [u] for F01, and 0% when the V is [i] for M03). Turning to V, its quality is an excellent predictor for TD position for F02 and M03, with on average 95% observations correctly classified as either [i] or [u]. The success rate is lower for speaker F01, especially for [u] (81% for [i] and 64% for [u]).

Table 1: Cross-validation (in %) of the linear determinant analysis with tongue displacement during the schwa.

Schwa	**F01**		**F02**		**M03**	
	[i]	**[u]**	**[i]**	**[u]**	**[i]**	**[u]**
[i]	59	38	72	73	0	8
[u]	41	62	28	27	100	92

Table 2: Cross-validation (in %) of the linear determinant analysis with tongue displacement during V.

V (i/u)	**F01**		**F02**		**M03**	
	[i]	**[u]**	**[i]**	**[u]**	**[i]**	**[u]**
[i]	81	36	97	6	95	5
[u]	19	64	3	94	5	95

Hence, TD position during V affects TD position during the schwa for speaker F01. This pattern holds when V is [i] for F02, and when V is [u] for M03. V quality has an effect on TD position during the V as well (less for speaker F01). The effect of V's quality on TD position during both the schwa and the V itself is further confirmed by four sets of linear mixed effects models, which examined TD horizontal and vertical displacement as a function of Vowel (levels: [i], [u]) and Unknown (levels: unknown consonant, unknown vowel) for the schwa and the V respectively. Item nested within speaker was treated as random effects. The analyses on the schwa revealed an effect of Vowel on the horizontal TD displacement (t= 2.81, p= 0.005), but not of Unknown. Neither factor was significant with respect to the vertical TD displacement. The analyses of the V showed an effect of Vowel on both the horizontal (t= 11.92, p< 0.001) and the vertical displacement (t= 3.28, p= 0.001). The factor Unknown was not significant.

4. DISCUSSION

Our results reveal two distinct speaker-dependent patterns. On the one hand, speaker F01 showed anticipatory coarticulatory effects on the first vowel of [əˈbVCɑ] that differentiated between schwas preceded by [i] and [u], while not clearly distinguishing between [i] and [u] during V, although V was known by the time it was produced. On the other hand, speakers F02 and M03, who clearly defined [i] and [u] during V, did not present differential anticipatory coarticulation during the schwa. F02 adopted an [i]-like and M03 an [u]-like TD position during the production of the schwa.

One possible interpretation of these patterns is that speakers adopted one of two different strategies for performing the task. Speaker F01 anticipated the following vowel and planned for it, showing respective anticipatory coarticulatory effects on the preceding vowel (schwa) and mixed articulatory target positions for the anticipated vowel (V). Speakers F02 and M03 treated the schwa as a preparatory stage for the upcoming planning and assumed a default articulatory configuration [cf. 9]. This configuration involved a TD position that was more similar to [i] for speaker F02 and to [u] for speaker M03. Another possibility is that speaker F02 was more often predisposed to [i] in the V position, and speaker M03 to [u], in which case the spatial profile of their schwas reflects anticipatory coarticulation driven by their respective predispositions. Both possible interpretations of speakers' strategies are further supported by the duration measures reported in the acoustic component of the current study [19], according to which speakers F02 and M03 had longer schwas than F01, who had longer bilabial stops in the second syllable instead. These measures also confirm that the speakers incorporated the missing segment successfully into the utterance.

These conclusions further suggest that anticipatory coarticulation depends on planning, thus supporting [18], and is not simply an artefact of inertia. Focusing on the articulatory aspect of coarticulation offers a promising window into planning, with possible theoretical and practical extensions [cf. 14, 16]. We are currently continuing our investigation by examining other tongue body positions (it is possible that some coarticulatory effects were not detected here because of the very posterior location of the TD sensor), more speakers, and more conditions. These include strings not involving a missing segment, and contrasts in production amplitude (i.e., loudly, when we expect coarticulation to be naturally perturbed, vs. normal levels).

Acknowledgments: This work was supported by NIH grant DC-002717 to Haskins Laboratories.

5. REFERENCES

[1] Beddor, P.S., Harnsberger, J.D., Lindemann, S. 2002. Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. 30, 591-627.

[2] Bell-Berti, F., Harris, K.S. 1981. A temporal model of speech production. 38, 9-20.

[3] Benguerel, A.-P., Cowan, H.A. 1974. Coarticulation of upper lip protrusion in French. 30, 41-55.

[4] Fowler, C.A. 2005. Parsing coarticulated speech in perception: Effects of coarticulation resistance. *J. Phonetics* 33, 199-213.

[5] Fowler, C.A., Brown, J.M. 1997. Intrinsic f0 differences in spoken and sung vowels and their perception by listeners. *Perc. & Psychophys.* 59, 729-738.

[6] Fowler, C.A., Smith, M. 1986. Speech perception as “vector analysis”: An approach to the problems of segmentation and invariance, in *Invariance and variability in speech processes*, J. Perkell and D. Klatt, Editors, Lawrence Erlbaum Associates: Hillsdale, NJ. p. 123-136.

[7] Furui, S., Deng, L., Gales, M., Ney, H., Tokuda, K. 2012. Fundamental technologies in modern speech recognition. *IEEE Signal Processing Magazine* 29(6), 16-17.

[8] Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychol. Rev.* 74, 431-461.

[9] Gick, B., Wilson, I., Kock, K., Cook, C. 2004. Language-specific articulatory settings: Evidence from inter-utterance rest position. *Phonetica* 61, 220-233.

[10] Liberman, A.M., Delattre, P.C., Cooper, F.S., Gerstman, L.J. 1954. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs* 68, 1-13.

[11] Liberman, A.M., Whalen, D.H. 2000. On the relation of speech to language. *Trends Cog. Sci.* 4, 187-196.

[12] Magen, H.S. 1997. The extent of vowel-to-vowel coarticulation in English. *J. Phonetics* 25, 187-205.

[13] Morgan, N., Wegmann, S., Cohen, J. 2013. What's Wrong With Automatic Speech Recognition (ASR) and How Can We Fix It? DTIC Document.

[14] Nallasamy, U., Metze, F., Schultz, T., Enhanced polyphone decision tree adaptation for accented speech recognition, in InterSpeech 2012. 2012: Portland, OR.

[15] Pardo, J.S., Fowler, C.A. 1997. Perceiving the causes of coarticulatory acoustic variation: Consonant voicing and vowel pitch. *Perc. & Psychophys.* 59, 1141-1152.

[16] Variani, E., Li, F., Hermansky, H. 2013. Multi-stream recognition of noisy speech with performance monitoring. in InterSpeech 2013. 2013: Lyon, France.

[17] Whalen, D.H. 1989. Vowel and consonant judgments are not independent when cued by the same information. *Perc. & Psychophys.* 46, 284-292.

[18] Whalen, D.H. 1990. Coarticulation is largely planned. *J. Phonetics* 18, 3-35.

[19] Whalen, D.H., Katsika, A., Tiede, M.K., King, H. in press. Acoustic measures of planned and unplanned coarticulation, in *the 18th International Congress of Phonetic Sciences,* 2015. 2015: Glasgow, UK.