ACOUSTIC MEASURES OF PLANNED AND UNPLANNED COARTICULATION

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

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

ABSTRACT

We present a replication of an experiment [23] where vowel-to-vowel coarticulation was found when an utterance was initiated before an intervening consonant's identity was known. English nonsense strings [əˈbVCɑ] were used. In one condition, the V was known but the C was not, with the reverse in the other condition; missing information was presented once phonation began. Results show anticipatory vowel effects on the final portion of the schwa (transitions into the stop) occurred only when the vowel was known ahead of time. Anticipatory effects of the V on the schwa's F2 were found throughout its duration when the V was known, but only for the speaker with the shortest V duration. Perseverative effects on the final vowel were similar in both conditions, as was consonant coarticulation. The implications of these results for planning are discussed, and the value of replication in the social sciences is emphasized.

Keywords: coarticulation; acoustics; planning; motor control; replication.

1. INTRODUCTION

Coarticulation, the influence of one segment on another, is the means by which speakers can overlap segments and increase the rate at which the ear can resolve individual sounds [12, 14]. Without the ability to overlap segments in time, the ear would not be able to discriminate sounds or judge their order, both critical for conveying a message. From a signal processing point of view, coarticulation is problematic because it makes analysing the speech signal a challenge that has yet to be fully met [9, 16]. While computer speech applications have had great success in recent years, difficulties still arise with changes in accent or dialect [17] or noisy conditions [21]. Extending the techniques currently in use to under-resourced languages is also a challenge [4], even though human are reported to learn all languages with equal facility. Unravelling the source of coarticulation can be expected to improve speech processing in these domains.

There is substantial perceptual evidence that human listeners do not have difficulty with coarticulation, but rather depend on it. Listeners parse each component into its various sources [5, 7, 19]. Thus an acoustic feature such as fundamental frequency (F0) can simultaneously inform decisions about vowel height and pitch [6], and duration can inform both a vowel quality and a consonant voicing judgment [22]. Such a perceptual strategy allows us to take advantage of inherent knowledge of the acoustic consequences of speech articulator movements.

Coarticulation is easiest to see in the formant transitions immediately after release of a consonant constriction [13], but segments that are relatively far apart can be affected as well [2, 3]. In particular, vowels can influence each other across intervening consonants [1, 15]. The extent to which this effect is an inertial response versus a planned change in articulation was studied acoustically in [23].

In that study, all but one segment of a brief, nonsense utterance (e.g., [əˈbibɑ]) was known to the speaker before speech was initiated via a text on a computer screen, with one letter missing. The utterance began with a schwa and the missing segment was either a consonant or a vowel of a subsequent syllable. Once phonation began, the missing segment's letter appeared on screen, and the speaker finished the utterance as naturally as possible. Anticipatory effects of the upcoming stressed vowel were seen on the schwa's F2, especially during the transitions into the stop, even if the effect crossed a segment (the consonant) whose identity was not known. Effects were reduced or eliminated when the vowel was not known at initiation. Perseverative effects on the third syllable were unaffected by this condition. The results were interpreted as showing that planning was responsible for the large temporal extent of the anticipatory effects.

The present experiment replicates Experiment 2 of that earlier study, in which English nonsense strings [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. In the most relevant situation, the end of the schwa was affected by the [i] or [u] only

when it was known at the outset of the utterance, even though this portion of the vowel could reasonably be seen as an inertial component of coarticulation. The onset of the schwa was affected by the (known) [i] or [u] for two of the three speakers as well in that condition, indicating substantial planning for at least some speakers.

2. EXPERIMENT

Our experiment followed the same procedure used in Experiment 2 of [23]. In addition to the acoustic signal, we recorded articulation via electromagnetic articulometry (EMA; WAVE, NDI). Only the acoustic results are reported here.

2.1.1. 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.

2.1.2. Stimuli

English nonsense strings [əˈbVCɑ] were used. The V was [i] or [u] while the C was [b] or [p].

2.1.3. Procedure

There were 8 stimuli, repeated (in random order) 20 times (F01) or 12 times (F02 and M03).

Custom software (*Marta*, Haskins Laboratories) was used to control stimulus presentation and audio recording. Utterances were digitized at 44.1 kHz. Each speaker was seated before a computer screen which displayed the stimulus at the beginning of a trial with one letter missing (for the vowel unknown condition, "UHB_BA;" for the consonant unknown, "UHBI_A"). Participants were instructed to begin production of the incomplete stimulus. Phonation amplitude was monitored continuously, and once it exceeded a target threshold the missing letter appeared (I or U in the first case, B or P in the second) and the speaker attempted to incorporate it as smoothly as possible into the utterance.

2.1.4. Analyses

Durations were determined from the acoustic waveform by hand measurement, segmenting the initial schwa, C1 duration, V duration (of 2nd syllable), C2 duration, and final vowel duration. Frequency measures (F0 and formants) were obtained for ten frames within each schwa: two frames at the beginning; one frame at the acoustic midpoint of the schwa; and at seven consecutive offsets going into the first stop closure (i.e., the transitions). 30 ms windows were used to resolve formants with centers of consecutive frames being separated by 10 ms. For short tokens, the middle frames would coincide with one of the frames of the transitions; this occurred 58% of the time, and the frames contributed to both measures. In 3% of the cases, the first two frames overlapped as well.

If the present results replicate Whalen [23], we expect to find:

- the second vowel (V) will affect the transitions from the schwa into the first stop only when it is known at initiation; specifically, we expect that the F2 transition of the schwa should be higher in the /əˈbiCɑ/ case than in the /əˈbuCɑ/ case, reflecting coarticulation of the more fronted tongue position required for the [i] (compared to the [u]).
- some speakers will show the vowel's influence at the onset of the schwa.
- the duration of the schwa will be greater when the vowel is unknown than when the consonant is.
- the final syllable will be unaffected by the known/unknown conditions.

3. RESULTS

Figure 1 shows the second formant of the schwa at the ten measurement points plotted separately for each of the speakers. Individual analyses were called for due to the small number of speakers and the lack of overlap in the target values of the schwa itself.

The vowel affects the final portions when it is known, but not otherwise. This can be seen in Fig. 1 by the separation of the C_unknown_i line from the others, with [i] coarticulating to make F2 higher in the schwa. In a repeated measures ANOVA done separately for each speaker with the factors Unknown (C or V) and Vowel (i or u), the final frame showed an effect of vowel only in the C-unknown condition for F01 and M03 ($F(1, 118) = 5.1136$, $p < 0.05$, $F(1, 76) = 12.679$, $p < 0.001$, respectively, for the interaction of Unknown and Vowel; the pairwise comparison of just the C-unknown condition was significant for both). For speaker F02, the interaction was marginal for the end point ($F(1, 56)$ 2.6587, $p = 0.11$) but appeared in the mid frame ($F(1, 57) = r.22$, $p < 0.05$, for the interaction).

Anticipatory effects were evident in the initial portion for F01, but not the other two speakers ($F(1, 117) = 7.42$, $p < 0.05$, for the interaction). The other two speakers showed no effect of the upcoming vowel in either condition. It is worth noting that the schwa duration was the shortest in the C unknown condition for F01 (115 ms) compared to the other two speakers (F02: 307 ms; M03: 233 ms).

Figure 1: F2 values for ten measurement points (10 ms frames) in schwa (see text). Speakers F01, F02 and M03 are shown separately.

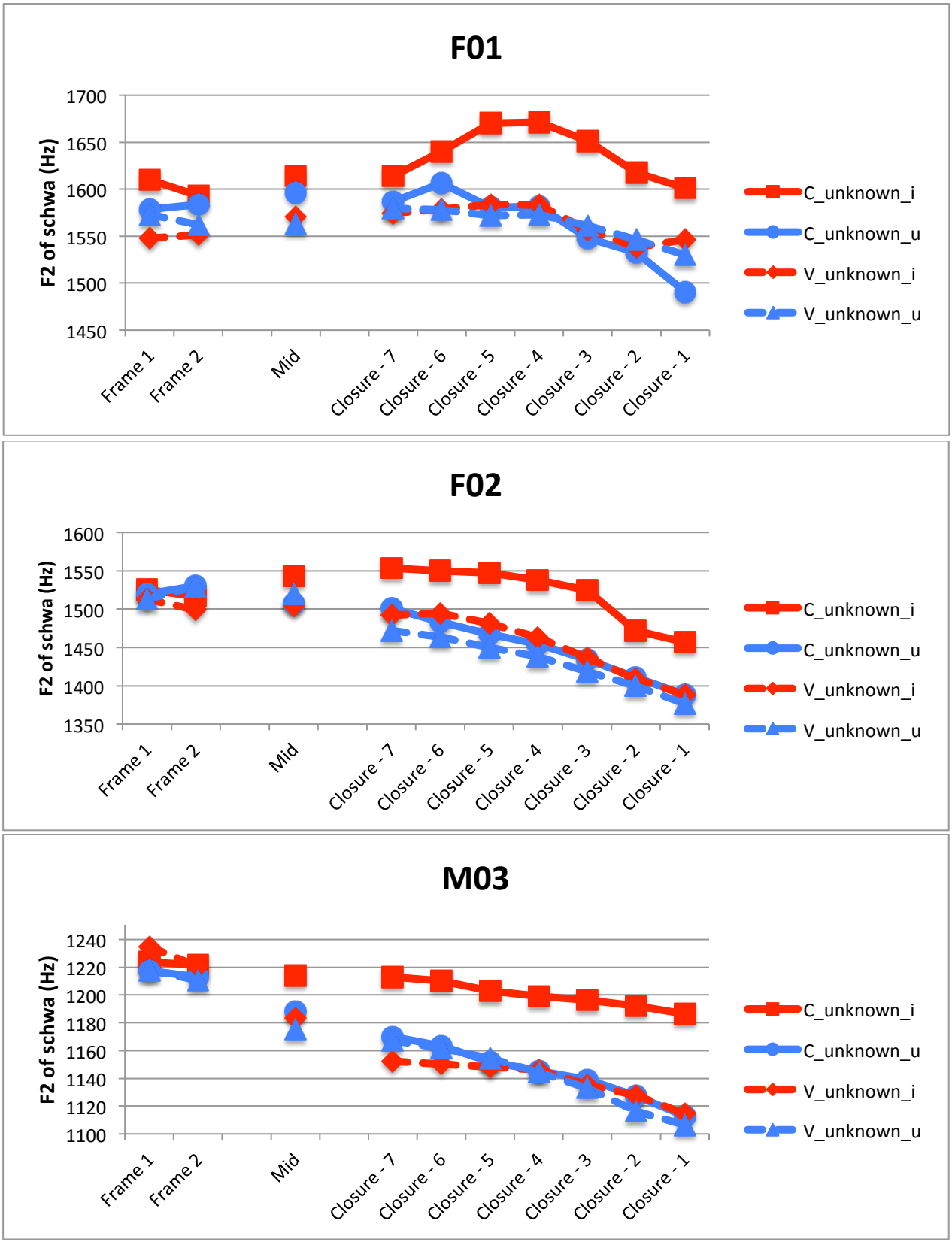
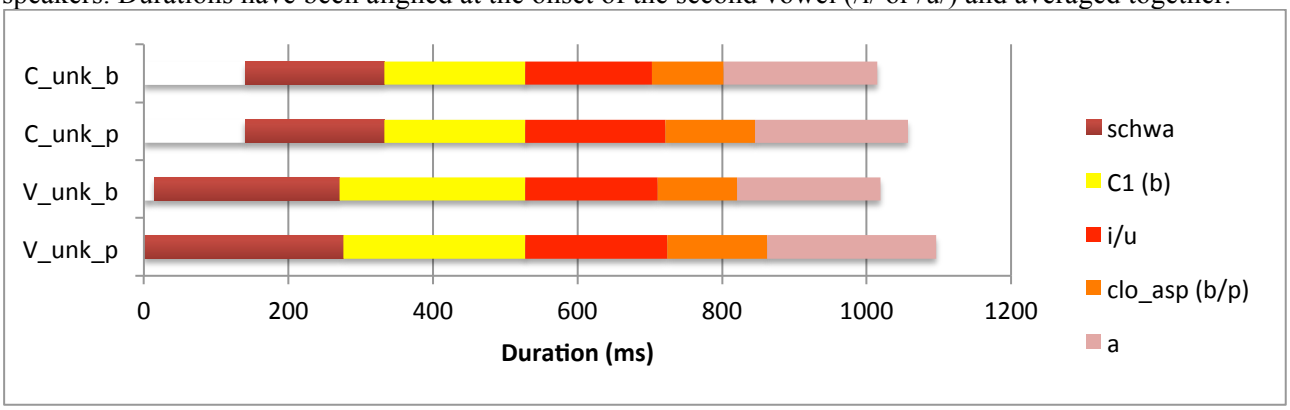


Figure 2: Durations of the components of the three syllables of [əˈbVCɑ], averaged across the three speakers. Durations have been aligned at the onset of the second vowel (/i/ or /u/) and averaged together.



The durations are shown in Figure 2. As was found previously, preparing for the unknown vowel extended the duration of both the schwa and the /b/ preceding the /i/ or /u/. The durations here were much longer than those in [23]. The third syllable's duration was also uninfluenced by condition (though syllables with /p/ were slightly longer), indicating that the changes needed for producing the unknown consonant were in place during the second syllable.

The final syllable's F2 was affected by the consonant (b/p), with the aspirated [p^h] having about 100 Hz higher onset than the unaspirated "b" ([p]). This has been reported previously [20]. Space limitations prevent a full display of these results, but it is clear that the third syllable was produced similarly in the two conditions, indicating successful integration of the missing elements.

4. DISCUSSION

The present results provide a replication of the earlier study in [23]. When the upcoming vowel is known, the higher F2 of /i/ can be seen in a higher F2 for the end of the schwa (transitions into the stop) compared to the other conditions. As in the earlier work, this was true for all three speakers, although F02's results were somewhat more complex. Similar effects were found for one speaker (F01) at the onset of the schwa. Although the earlier work has found this effect for two out of three speakers, the durations of the schwa for F02 and M03 were much longer than those of the earlier speakers and of F01. The schwa was longer in the condition where the vowel of the next syllable was unknown than in the condition where the consonant of the third syllable was not known. Condition did not affect the acoustics of the third syllable. Thus the speakers were relatively successful at incorporating the segment that was unknown at utterance initiation into the remainder of the utterance.

The results support the main conclusions of [23]. Early vowel-to-vowel coarticulation shows a dependence on planning, especially during the transitions into the first stop. The onset effect, though present for two speakers in the earlier work, was present only for one of three here. Besides the duration difference mentioned earlier, it is worth noting that the protocol was changed (to increase training and reduce the overall length) after F01 participated, and she produced twice as many tokens as the other talkers. It is possible that the number collected from the later talkers was not sufficient to allow the effect to appear through the normal range of production variability and/or the known problems with formant analysis [11].

Further issues remain to be explored in this paradigm. The changes in the articulators are of particular interest and are under study, with preliminary results corroborating the patterns presented here [10]. The fact that the /u/-influenced formant values for schwa are still higher than the uncoarticulated ones, rather than the predicted lower values, requires further explanation.

The need for replication has been a basis of science for centuries, but recent concerns over the lack of replication have made it clear that the social sciences in particular are lacking in this regard [8, 18, 25]. Even though many results in the phonetic literature are implicitly replicated, fewer are replicated overtly. Further, there is a clear effect of the number of speakers analyzed and the convergence of results [24]. The present study is a step in the direction of replication, but the additional complexity of articulatory measures ensures that the total number of speakers will remain limited. We hope to extend beyond the current number in order to more fully understand the role of planning in speech motor control as exemplified by coarticulation.

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

7. 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. *J. Phonetics* 30, 591-627.

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

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

[4] Besacier, L., Barnard, E., Karpov, A., Schultz, T. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Comm.* 56, 85-100.

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

[6] 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.

[7] 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.

[8] Francis, G. 2012. Publication bias and the failure of replication in experimental psychology. *Psychonom. Bull. Rev.* 19, 975-991.

[9] 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.

[10] Katsika, A., Whalen, D.H., Tiede, M.K., King, H. in press. Articulatory measures of planned and unplanned coarticulation. in 18th International Congress of Phonetic Sciences. Glasgow.

[11] Klatt, D.H. 1986. Representation of the first formant in speech recognition and in models of the auditory periphery, in *Proceedings of the Montreal satellite symposium on speech recognition, 12th International Congress on Acoustics*, P. Mermelstein, Editor, Canadian Acoustical Society: Montreal. p. 5-7.

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

[13] 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.

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

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

[16] Morgan, N., Wegmann, S., Cohen, J. 2013. What's wrong with automatic speech recognition (ASR) and how can we fix it? DTIC Document.

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

[18] Nosek, B.A., Lakens, D. 2014. Registered Reports. *Social Psych.* 45(3), 137-141.

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

[20] Repp, B.H., Lin, H.-B. 1987. Difference in second-formant transitions between aspirated and unaspirated stop consonants preceding [a]. *Language and Speech* 30, 115-129.

[21] Variani, E., Li, F., Hermansky, H. 2013. Multi-stream recognition of noisy speech with performance monitoring. in INTERSPEECH 2013. Lyon.

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

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

[24] Whalen, D.H., Levitt, A.G. 1995. The universality of intrinsic F0 of vowels. *J. Phonetics* 23, 349-366.

[25] Wolfe, J. 2013. Registered Reports and Replications in Attention, Perception, & Psychophysics. *Atten. Percept. Psychophys.* 75, 781-783.