A psycholinguistic method for measuring coarticulation in child and adult speech

Phil J. Howson¹ · Jeffrey E. Kallay¹ · Melissa A. Redford¹

Published online: 1 September 2020
© The Psychonomic Society, Inc. 2020

Abstract

A perceiver's ability to accurately predict target sounds in a forward-gated AV speech task indexes the strength and scope of anticipatory coarticulation in adult speech (Redford et al., *JASA*, 144, 2447–2461, 2018). This suggests a perception-based method for studying coarticulation in populations who may poorly tolerate the more invasive or restrictive techniques used to measure speech movements directly. But the use of perception to measure production begs the question of confounding influences on perceiver performance and thus on the reliability and generalizability of the proposed method. The present study was therefore designed to test whether a gated AV speech method for measuring coarticulation provides reliable results across different study populations (child versus adult), different task environments (in-lab versus online), and different coarticulatory directions (forward/anticipatory versus backward/carryover). The results indicated excellent measurement reliability across age groups in the forward/anticipatory measurement direction, though more perceivers are needed to achieve the same levels of agreement and consistency when the task is completed online. Accuracy was lower in the backward/carryover direction, and although agreement and consistency were still reasonably high across perceivers, the effect of age group differed between the laboratory and online environments, suggesting measurement error in one or both environments. Overall, the results support using in-lab or online perceptual judgments to measure anticipatory coarticulation in developmental studies of speech production. Further validation study is needed before the method can be extended to measure carryover coarticulation.

Keywords Gating paradigm · AV speech perception · Speech production · Online perception task

In previous work, we demonstrated that aggregated perceiver responses in a forward-gated audiovisual (AV) speech prediction task yields a measure of coarticulatory onset and strength in adult speech that conforms with literature-based expectations and direct video-based measurement of orofacial movement (Redford, Kallay, Bogdanov, & Vatikiotis-Bateson, 2018). As in other production studies, the gated AV speech method for measuring coarticulation requires that speakers are recorded while producing several repetitions of sentences that contain contrasts meant to elicit coarticulatory effects of interest. The sentences are excised from the AV recordings and gated to answer experimental questions about the scope and strength of long-distance coarticulation. The perceiver's task is to guess the targeted contrast when presented with gated stimuli, blocked by speaker. The measure of coarticulation is

bias-corrected prediction accuracy, aggregated across perceivers by gate and speaker. Comparison with acoustic measurement and perceiver performance on gated audio-only speech in the Redford et al. study indicated that AV speech perception also provides a more accurate measure of coarticulatory onset than speech acoustics alone. Such a result is not surprising, since perceivers can use visual cues in AV speech to detect anticipatory movements that are difficult to hear or to measure. Importantly, these cues are not limited to the speaker's lips. The coordinated movements of a speaker's tongue, lips, and jaw deform the soft tissues of the face so completely that its summary measurement predicts speech acoustics with the same level of accuracy as direct kinematic measurement of the speech articulators (Yehia, Rubin, & Vatikiotis-Bateson, 1998).

The tight correlation between facial movement and acoustics is encoded as implicit knowledge, built up from our vast experience as speakers and listeners of language (Campbell, 2008). This knowledge allows a perceiver to attend to all relevant speech movements in a flexible manner to infer speech goals. This is important because articulatory phonetic

✉ Melissa A. Redford
redford@uoregon.edu

¹ Department of Linguistics, 1290 University of Oregon,
Eugene, OR 97403, USA

studies amply demonstrate significant individual differences in speech sound production (see, e.g., Westbury, Hashi, & Lindstrom, 1998; Baum & McFarland, 2000; Derrick & Gick, 2011) and significant within-speaker differences in how a sound is realized depending on context (i.e., motor equivalence). Tapping into a perceiver's implicit knowledge to measure coarticulation abstracts away from these sources of variation, which is useful for the psychological study of speech production where the speaker's intended goal is of greater interest than articulatory detail.

Perceiver's implicit knowledge also provides a normalized measure of production. Studies by Jordan and colleagues demonstrate that perceiver performance in AV speech perception tasks is insensitive to the size of the speaker's image on the screen (Jordan & Sergeant, 1998), to the speaker's orientation with respect to the camera (Jordan et al., 1997; Jordan & Thomas, 2001), and, surprisingly, even to whether or not the speaker's face is upside down in the frame (Jordan & Bevan, 1997). A perceiver's ability to automatically normalize for size and perspective implies that coarticulation can be measured perceptually in gated AV speech without restricting a speaker's natural non-speech movements. This is especially useful for studying coarticulation in populations who may poorly tolerate the more invasive or restrictive techniques that are used to directly measure speech movements. In fact, it is our specific interest in one such population, children, that led us to validate the gated AV speech method for measuring coarticulation in the first place.

Of course, implicit knowledge about speech also renders perceivers susceptible to illusions that may undermine measurement reliability. The McGurk effect provides a well-known example of an AV speech illusion (e.g., McGurk & MacDonald, 1976). To achieve this illusion, perceivers are presented auditorily with one syllable (e.g., *ba*) while watching someone articulate a different syllable (e.g., *ga*). Perceivers report hearing a third syllable, which is intermediate (in some sense) to the presented syllables; for example, if the auditory stimulus has the bilabial onset, /b/, and the other has the velar onset, /g/, then perceivers will report hearing a syllable with an alveolar onset, /d/. This effect depends on the perceivers' implicit knowledge that the visual and auditory speech are always congruent (Campbell, 2008). Put another way, a perceiver's lack of experience with incongruent visual and auditory cues forces the perceiver to resolve the AV speech signal so that it is consistent with their implicit knowledge of AV speech even while it is inconsistent with the real AV speech stimulus. Thus, a lack of experience with particular speech processing conditions or speaker groups could disrupt the reliability of perception-based measurement of speech production that relies on AV speech.

In sum, a perceptual measure of coarticulation based on gated AV speech holds significant promise for the psychological study of speech production, especially in vulnerable

populations. But the use of perception to measure production relies on implicit speech knowledge, which begs the question of confounding influences on perceiver performance. The current study was therefore designed to test measurement reliability across several conditions of interest, including different speaker populations, different experimental manipulations of speech, and different task settings.

Current study

As noted, the gated AV speech method for measuring coarticulation was proposed to study the development of speech production. But many of the college-aged adult perceivers we rely on to provide responses for our measure may lack experience interacting with children. It is possible that these perceivers are less adept at identifying cues to coarticulation in children's speech than they would be at identifying such cues in speech produced by other college-aged adults. After all, direct measurement studies strongly suggest that children adopt different coarticulatory strategies than adults (Zharkova, Hewlet, & Hardcastle, 2008; 2011; Noiray, Ménard, & Iskarous, 2013; Noiray, Abakarova, Rubertus, Kruger, & Tiede, 2018; Rubertus & Noiray, 2018). For example, Noiray et al. (2018) used ultrasound to study lingual coarticulation in monolingual German-speaking children between the ages of 3 and 7 years old. The authors found that coarticulatory effects across syllable boundaries were higher in children than in adults. Children's articulation of intervocalic consonants (/b/, /d/, /g/) was also more strongly influenced by an upcoming vowel than adults' articulation. Importantly, the findings of differences between child and adult speech were not restricted to the youngest children: even 7-year-old children were found to differ from adults in the articulation of some sequences. Such differences between child and adult speech likely reflect differences in articulatory coordination due to the slow maturation of speech motor skills (Smith & Zelaznik, 2004). If college-aged perceivers are not sensitive to these maturational differences, their prediction of sounds based on coarticulatory cues may be more variable (and therefore less reliable) in children's AV speech than in adults' AV speech. The current study tests this prediction.

Just as a lack of experience with children's speech may adversely impact the ability of college-aged adults to reliably predict sounds from the coarticulatory cues provided in AV speech, so too might perceiver performance be affected by the naturalness of the prediction task. The forward prediction task that we validated in Redford et al. (2018) is a reasonably natural task in that there is ample psycholinguistic evidence to suggest that listeners make use of prediction to facilitate speech processing and language comprehension (see, e.g., Federmeier, 2007; Pickering & Garrod, 2007), including the

use of anticipatory cues to speech sound articulation (Salverda, Kleinschmidt, & Tanenhaus, 2014). Forward prediction also aligns perfectly with the goal of measuring anticipatory coarticulation. But, to paraphrase Dell and colleagues (Dell, Schwartz, Martin, Saffran, & Gagnon, 1997), speech production is not only about predicting the future; it is also about inhibiting the past. Thus, we wondered: can we use perceiver performance in a gated AV speech task to measure carryover (i.e., left-to-right or perseveratory) coarticulation?

The obvious way to measure carryover coarticulation in AV speech is by backward-gating the speech and asking perceivers to guess what sound occurred earlier in the speech stream. This task is clearly less natural than the forward prediction task, though maybe not completely unnatural. Listeners have experience recovering occluded speech sounds (see Warren, 1970), of which backward-gated speech might be an extreme example. But perceptual recovery uses implicit knowledge about intended speech sounds imparted by the word or phrase within which the occluded sound appears. If carryover effects are not encoded as part of implicit speech knowledge, then it might be difficult for the perceiver to reconstitute missing sounds. This possibility is suggested by the assumption that carryover effects are epiphenomenal rather than planned (see, e.g., Whalen, 1990). If this assumption is correct, then perceivers may have little motivation for attending to carryover cues in normal speech processing, as these would provide little information about a speaker's intentions. Absent attention to carryover coarticulation, perceivers may not have adequate knowledge to reliably use such cues in a backward speech prediction task. Accordingly, perceiver performance in the backward-gated AV speech task would provide a less reliable measure of coarticulation than performance in the forward-gated version. The current study also tests this prediction.

Given that perceiver attention is critical for collecting reliable data during the execution of a task, the current study also tested for an effect of environment on perceiver performance. More specifically, we compare the reliability of data collected in the laboratory to data collected online. Although previous studies have found that online data is at least as reliable as laboratory data (Goodman, Cryder, & Cheema, 2013; Shapiro, Chandler, & Mueller, 2013; Arditte, Çek, Shaw, and Timpano, 2016; Hauser & Schwarz, 2016; Miller, Crowe, Weiss, Maples-Keller, & Lynam, 2017), it is important to verify online data reliability for a perceptual task such as ours, since the lack of control over the online experience may present a significant confound. If we find that online judgments of gated AV speech are as reliable as laboratory judgments, then our method can be scaled up to measure coarticulation in speech produced by greater numbers of speakers than is feasible when the measurements are acquired in the laboratory.

Experiment 1

Experiment 1 tested whether the forward-gated AV speech prediction task provides a reliable measure of anticipatory coarticulation in children's speech. Vowel-to-vowel coarticulatory effects have been shown to span unstressed syllables in adult and child speech (e.g., Magen, 1997; Grosvald, 2009; Kallay & Redford, 2018; Rubertus & Noiray, 2018). For example, Kallay and Redford (2018) investigated both anticipatory and carryover effects on the production of an unstressed determiner in simple subject-verb-*the*-object sentences elicited from 5-year-olds and college-aged adults (e.g., Maddy packs *the* gak). The stressed vowels in the verb and object noun were varied to be either /æ/ or /ou/ in order to examine the effect of the three-way contrast (= jaw height, tongue advancement, and lip rounding) on the production of the unstressed determiner vowel (= schwa). Coarticulatory strength was measured by calculating the Euclidian distance in acoustic (F1 × F2) space in the anticipatory and carryover contexts. Analyses indicated anticipatory effects on schwa across age groups, with more reliable effects in adults' speech; carryover effects were also found, albeit only in children's speech. Anticipatory effects of a singleton consonantal onset (/b/, /g/, or /s/) on schwa production were also investigated. No difference between age groups emerged in the results, but there were strong effects of consonant. An inspection of F1 values suggested that schwa before /b/ was more open than before /s/; F1 values for /g/ were intermediate to /b/ and /s/. An inspection of F2 values suggested that the tongue was more retracted for /b/ than for either /s/ or /g/.

A subset of AV recorded sentences produced by 12 speakers from the Kallay and Redford (2018) study were used in the current study. The forward-gated version of the prediction task used the frame sentence, *Maddy packs the _____*, where the object noun was either *gak* or *goat*. These sentences were gated from the midpoint of the /æ/ vowel in the verb to the midpoint of the target /æ/ or /ou/ vowel across the unstressed determiner (*the*). The gated sentences served as stimuli that were blocked by speaker and presented to college-aged perceivers in random order. The perceiver's task was to guess whether the upcoming noun rhymed with "ack" or "oat." One group of perceivers completed the task in the laboratory, another completed the task online. Predictions based on prior work were that accuracy would be higher at gates closer to the target noun vowel than further away, and already above chance at the determiner onset (Redford et al., 2018). We also expected an effect of age group on prediction accuracy based either on age-related differences in the temporal extent and strength of anticipatory coarticulation and/or college-aged adult perceivers' lack of experience with children's speech. An effect of environment (laboratory versus online) on accuracy would suggest that perceivers' ability to detect coarticulatory cues varies with the task setting. An interaction

between age and environment on accuracy would suggest a need for caution in interpreting the results with respect to age-related differences in coarticulation. Perceiver agreement and consistency were expected to be equally high across age groups and environments, provided that the gated AV speech method is a reliable measure of anticipatory coarticulation and perceivers are able to detect cues to anticipatory coarticulation in children's speech. Detailed methods and results follow.

Method

Participants

Speakers Speakers were six 5-year-old children (5 yrs. 8 mos. to 6 yrs. 0 mos., with a mean age of 5 yrs. 10 mos.) and six college-aged adults who were also University of Oregon students. Like the college-aged adults, children were native American-English speakers living in the Eugene-Springfield area. Half of the children and adults were female. All speakers were screened for normal hearing; none had a history of speech-language therapy. All children also performed within the normal range on the Peabody Picture Vocabulary Test (PPVT-4; Dunn & Dunn, 2007) and on the single-word subtest of the Goldman-Fristoe Test of Articulation (GTFA-2; Goldman & Fristoe, 1999). Families and adult speakers were financially compensated for their time. Children also earned a small prize when they completed the study.

Perceivers Perceivers were 40 college-aged adults (18–23 years old). Twenty adults were assigned to the laboratory condition and 20 to the online condition. The perceivers had no self-reported history of speech or hearing disorders. All perceivers were recruited through the Linguistics and Psychology human subjects pool at the University of Oregon. They were compensated with course credit for their participation. Participants selected from a wide range of studies (over 50) to complete course credit. These studies were advertised by arbitrary names chosen from a list of composer names. Thus, the 40 students who selected our study for course credit did so blind to the type of experiment they were signing up for.

Stimuli

Materials As noted, the sentences used in the present study were drawn from a larger set elicited for an acoustic-phonetic study on anticipatory and carryover coarticulatory effects on determiner production in child and adult speech (see Kallay & Redford, 2018). The sentences took the form, *Maddy packs the TARGET* or *Maddy pokes the TARGET* where the target was a picturable object noun with an /æ/ or /ou/ vowel nucleus. Noun onsets were varied (/b, s, g/) to create six picturable words with contrasting rhymes: *bat* vs. *boat*, *sack* vs. *soap*, *gak* vs. *goat*.

Elicitation Sentences were elicited from child and adult speakers in the following way. The speaker sat across from an experimenter in a quiet observation room in front of a blue screen, facing a Panasonic HC-V770 audio-video camcorder. Lighting was provided by two Genaray SP-E-240B SpectroLED Essential 240 Bi-Color LED lights. The speaker wore a hat with a Shure ULX wireless microphone. The wireless receiver input directly to the camcorder. The experimenter first introduced the speaker to a set of cards with pictures that corresponded to the six object nouns to be elicited in the *pack* frame sentence. A picture of a spiky green cartoon creature was used for the target word *gak*¹ because a picture of the blob-like substance that is *gak*¹ was deemed less specific (i.e., less imageable) and less fun to name than the cartoon creature. Sentence elicitations were randomized by shuffling the object noun pictures within a block defined by the frame sentence. Each sentence was elicited once per block. There were five elicitation blocks per frame sentence. During the first few blocks, experimenters would produce the whole target sentence, introduce a pause by counting to three, and then ask the speaker to produce the sentence. Once the speaker knew what sentence to produce, the experimenter would simply provide the verb frame at the beginning of a block and then show each picture card to elicit the appropriate sentence. If the experimenter detected a disfluency or error during production, the card was placed to the back of the stack and the sentence re-elicited within the same block.

Gates Only sentences with *gak* and *goat* were used in the present experiment so that perceivers could judge sentences produced by all 12 speakers in a reasonable amount of time (< 45 minutes). Sentences with /g/-onset nouns were chosen in preference to those with /b/-onset nouns to allow for lip rounding as a potential cue to upcoming vowel identity. Sentences with /g/-onset nouns were also preferred to those with /s/-onset nouns, since the acoustic results from Kallay and Redford (2018) suggested that /g/-onsets may allow for more degrees of freedom in jaw opening during schwa production than /s/-onsets. Three good productions of *Maddy packs the gak* and *Maddy packs the goat* were identified for each speaker, starting with the second repetition. Good sentence productions were defined as those spoken under a single intonation contour without coughs, laughs, pauses, or other disfluencies. Once identified, the sentences were extracted from the AV recordings and cut using Lightworks Pro to create gated stimuli that preserved all speech information up to the cut. Cuts were made at segmental landmarks with reference to acoustic and kinematic cues to articulation. The cuts,

¹ Since *gak* is a real word, it was easily pronounced as such when applied to the spiky creature. Also, the phonestheme “-ack” (as in *whack*, *crack*, *smack*) seemed to us to be more consistent with a spiky creature than with a picture of a blob.

precise to within a 33-ms interval (recording speed was 30 frames per second), were made at the mid-point of the vowel in *packs* (Gate 1), the release of /k/ and onset of /s/ frication (Gate 2), the consonantal onset of the determiner (Gate 3), midpoint of the determiner vowel (Gate 4), consonantal closure during /g/ articulation (Gate 5), and midpoint of the target vowel (Gate 6). Figure 1 below shows the final video frame for the six gated stimuli derived from a single child's production of the sentence *Maddy packs the goat*. Each sentence repetition was cut in the same way. This resulted in 36 stimuli per speaker (2 sentences $\times$ 3 repetitions $\times$ 6 cuts).

Procedure

Task Participants were informed that their perceptual judgments would be used to measure long-distance coarticulation. The concept of coarticulation was introduced with the example of the different articulations of the initial fricative in *stroop* versus *street*. Participants' attention was directed to the kinesthetic and visual differences in the /s/ of the two words, with the contrast between lip spreading versus rounding made especially salient. This example likely provided participants with a significant cue to coarticulation in the present study (i.e., un/rounding), but—more importantly—it helped to orient them to the specific task at hand, namely, making judgments about speech based on coarticulatory information. Recall that speech movements deform the entire face, not just the perioral region: tongue advancement is directly visible in the release of an occlusion and in any articulations that involve the teeth; tongue postures, and especially tongue retraction, can otherwise be inferred from jaw height and from soft tissue deformation of the cheek and neck (Yehia et al., 1998).

Once participants understood the concept of coarticulation, they were told that they would be seeing snippets of video where the speaker was producing either the sentence, *Maddy packs the gak*, or the sentence, *Maddy packs the goat*. They were instructed to guess, based on the snippet, whether the speaker intended the final noun *gak* or *goat* by selecting an “ack” or “oat” response button, respectively. This button was displayed on the video screen with the final frame of the gated

sentence. These buttons remained on the screen until the participant selected a response. Trials were blocked by speaker and randomized within each block. The order of the blocks was also randomized.

The online experiment was conducted using the platform Testable (www.testable.org). This platform allows presentation of visual and/or audio stimuli with customization for inter-trial pause durations, inter-stimulus pause durations, participant answer selections, and randomization and blocking capabilities, among other customization options. We designed the online experiment to mimic the laboratory settings as closely as possible. The blocking, randomization, and buttons were all the same. The only task differences between the online and laboratory settings were that (1) instructions were delivered in text in the online setting, and (2) participants were asked to calibrate their video using a credit card so that the size would be a constant 1280×720 pixels. In the laboratory setting, instructions were given orally and the video size was always 1920×1080 pixels. In the laboratory setting, all participants used a set of AKG K240 MKII headphones, while the online participants used whatever headphones they had available. Calibration data were recorded in the online condition. These data indicated that perceivers generally followed the instructions and video size was controlled. What little variability existed did not correlate with performance: Pearson's $r(20) = .051$, $p = .83$. No additional information was recorded or elicited in the online setting regarding setup (e.g., device type, internet stability, headphone type).

Statistical analyses

Accuracy Correct responses were summed within each perceiver across repetitions within gate and speaker. Response bias was corrected by calculating the rand index accuracy (Rand, 1971) as the sum of true positives and negatives divided by the sum of all response types, including false positives and negatives (i.e., $\text{Accuracy} = \text{TP} + \text{TN} / \text{TP} + \text{FP} + \text{TN} + \text{FN}$). In this study, the stressed vowel /oo/ in target position required deviating from the articulatory posture for the preceding stressed vowel /æ/ (in the verb *pack*). Thus, an *oat*

Fig. 1 The final frame from each of six stimuli derived from a single production of the sentence “Maddy packs the goat” by a 5-year-old speaker. Gates were placed at the segmental landmarks shown, beginning with the verb, using acoustic and kinematic cues in the audiovisual recording

response to a gated *goat* sentence was treated as a true positive (TP); an *oat* response to a gated *gak* sentence was treated a false positive (FP); an *ack* response to a gated *gak* sentence was therefore the true negative (TN); and an *ack* response to a gated *goat* sentence was the false negative (FN). An accuracy of 1 equals perfect performance in the prediction task. An accuracy of .5 equals chance performance.

Accuracy coding resulted in 72 measures per perceiver (12 speakers $\times$ 6 gates). A mixed-effects linear model was used to test for the fixed effects of the speaker's age group (adult vs. child) and sentence gate (6 levels) on these accuracy measures (lme4 package; Bates, Mächler, Bolker, & Walker, 2015). The random effects of perceiver and speaker were controlled by including a random slope for every level of the interaction between perceiver and speaker in the model. The analyses were first conducted on data collected in the laboratory setting and then on data collected in the online setting. An effect of setting on performance was tested by including condition (laboratory vs. online) as a fixed effect in a subsequent overall analysis. Significant effects were tested with the R anova function (R Core Team, 2018). Planned comparisons between gates were done using the emmeans package (Lenth, 2018). Figures were created using ggplot2 (Wickham, 2016).

Reliability Two-way random-effects intraclass correlation coefficient analyses (ICC; Koo & Li, 2016) for average measures were performed to examine agreement and consistency among perceivers by age group and condition using the irr package version 0.84 (Gamer, Lemon, Fellows, & Singh, 2012) in R (R Core Team, 2018). Agreement refers to whether different raters had the same rating for the same stimuli. A higher rating indicates more agreement among raters. In the context of this experiment, high agreement means that perceivers gave similar ratings for the same speakers at the same gates. Consistency refers to whether an individual rater gave the same rating to repeated presentations of the same stimulus type. Higher consistency means that a rater was more consistent when rating repeated presentations. In the context of this experiment, high consistency reflects highly consistent ratings at the same gate for the same speaker. The interpretation of the ICC values as poor ($ICC < 0.50$), moderate ($0.50 < ICC < 0.75$), good ($0.75 < ICC < 0.90$), or excellent ($ICC > 0.90$) was based on the suggestions of Koo & Li (2016). The ICC package version 2.3.0 (Wolak, Fairbairn, & Paulsen, 2012) was used to determine the number of perceivers needed to achieve high agreement and consistency given effort, which is equal to number of participants by the number of observations per participant. In these calculations, the desired ICC value was specified with different confidence intervals and the number of observations held constant, leaving perceiver number free to vary.

Acoustics Kallay and Redford's (2018) acoustic study showed coarticulatory effects of the subsequent vowel on the unstressed vowel of the determiner (i.e., schwa), albeit in a larger sample of speakers (11 children and 9 adults) and across more different sentence types. Here, acoustic information for the sentences under investigation is presented, again for schwa in the determiner. Unlike in Kallay and Redford (2018), the current analyses are based on formant dynamics. F1, F2, and F3 values were extracted at 10 equally spaced intervals across the entire duration of schwa. Interpolated time course formant dynamics were compared as a function of the following noun, *gak* and *goat*, within each speaker age group. Comparisons were performed with best fit formant contours for F1, F2, and F3 with smoothing spline (SS) ANOVA (Davidson, 2006) models in R (R Core Team, 2018).

Results

Schwa formant dynamics

We begin with an analysis of adults' and children's schwa production as a function of the following stressed vowel to provide an acoustic context for the forward-gated AV speech results. Figure 2 presents SS-ANOVA plots of the F1, F2, and F3 trajectories in schwa of the determiner before "gak" versus before "goat" in the adults' (left) and children's (right) speech. The plots indicate that the anticipatory vowel-to-vowel coarticulatory effect was limited to F3 differences in adults' speech and to F2 differences in children's speech for the sentences under consideration in Experiment 1.

Given that F3 values were higher before /ou/ than before /æ/ in adults' speech, the coarticulatory effect on schwa was likely due to an anticipatory upper pharyngeal constriction for /ou/ (Stevens & Keyser, 1989) rather than to anticipatory lip rounding, which would lower F3. Similarly, the F2 difference in children's production of schwa may also suggest anticipatory tongue posturing for the following vowel (i.e., greater retraction for /ou/), but the lower F2 in the /ou/ context could also be due to anticipatory lip rounding. Note that the coarticulatory effects were not particularly stronger or weaker as a function of age group; they were simply different.

Laboratory setting

Turning now to the results from the forward-gated AV speech task, the analysis of perceiver performance in the laboratory setting revealed a main effect of sentence gate on prediction accuracy [$F(5, 1190) = 289.29, p < .001$], but no effect of the speaker's age group ($p = .515$) and no interaction between the factors ($p = .708$). The model R^2 was .48. The significant effect of gate, shown in Fig. 2, indicates that perceivers' ability to predict the rhyme of the upcoming noun increased with proximity to the target stressed vowel. The results shown in

Fig. 2 SS-ANOVA plots for the schwa in “the” produced by the 6 adult (left) and 6 child (right) speakers as a function of the following noun: formants for schwa before “gak” are shown in black; those for schwa before “goat” are shown in gray. Dotted lines indicate 95% confidence intervals

the figure also indicate that performance was at above chance levels from the onset of the determiner onward. Percent accuracy by gate, from the stressed vowel in the main verb through to the target stressed vowel in the object noun, was as follows (standard deviations in parentheses): 50% (13.6); 51.7% (14); 55.3% (17.3); 64.7% (20.2); 74.3% (19.7); 95.6% (12.5).

Mean comparisons, with alpha corrected for multiple comparisons, confirmed that prediction accuracy was lower when the sentence was gated at verb offset (“k/s” in Fig. 3) than when it was gated at determiner onset (“ð” in Fig. 3), which was in turn lower than prediction accuracy when the sentence was gated at the determiner vowel (“ə” in Fig. 3) and so on. Overall, the results are consistent with articulatory posturing for the target vowel as early as the onset of the determiner in both child and adult speech.

The ICC analysis revealed excellent reliability with coefficient values of greater than .90 across the different age groups. Table 1 summarizes the results for agreement and consistency. It also presents the result of *F*-tests for significance by age group for the analysis on average measures.

Online setting

Analysis of perceiver performance in the online setting revealed a similar main effect of sentence gate on prediction accuracy [$F(5, 1190) = 123.24, p < .001$], and a similar absence of an age group effect or interaction between gate and age group. The model R^2 was .28. The significant effect of gate, shown in Fig. 4, was as before: prediction accuracy increased with proximity to the target stressed vowel. Percent accuracy by gate, from the stressed vowel in the main verb

through to the target stressed vowel in the object noun, was as follows (standard deviation in parentheses): 51.9% (17.4); 50.8% (17.1); 55.1% (18.6); 60.4% (19.6); 68.1% (21.2); 85.6% (22.5). Thus, performance was again at above chance

Fig. 3 Perceiver performance in the forward prediction task is shown as a function of the speakers’ age (i.e., group: adult versus child) and sentence gate at the segmental landmarks indicated, where “æ” indicates the midpoint of the stressed vowel in “packs” and “T” the midpoint of the target vowel (/æ/ or /ou/) in the object noun. Whiskers indicate the 95% confidence intervals, and the dotted line indicates chance performance. The data are from perceivers who completed the task in a laboratory setting

Table 1 Agreement and consistency of perceiver performance in the laboratory condition

		Intraclass correlation	95% Confidence intervals		<i>F</i> -test	
			Lower bound	Upper bound	<i>F</i>	<i>p</i> -value
<i>Agreement</i>	<i>Child</i>	0.956	0.932	0.974	23.7	< .001
	<i>Adult</i>	0.960	0.939	0.977	26	< .001
<i>Consistency</i>	<i>Child</i>	0.958	0.935	0.975	23.7	< .001
	<i>Adult</i>	0.962	0.941	0.978	26	< .001

levels when sentences were gated at the determiner onset (Gate 4; “ð” in Fig. 4). Nonetheless, stringent tests for significant differences between means indicated that performance at this gate overlapped with performance on sentences gated at the midpoint of the preceding verb (Gate 1; “æ” in Fig. 4). Yet, the mean difference in prediction accuracy between the first gate (i.e., midpoint of verb) and Gate 5, the determiner vowel gate (“ə” in Fig. 4), was significant ($p < .001$). All other differences were also significant, save the difference between the first and second gates (= verb offset; “k/s” in Fig. 4) and the second and third gates (i.e., verb offset versus determiner onset).

Overall, the online accuracy results paralleled those obtained in a laboratory setting, but—as the R^2 indicates—the results were somewhat weaker. The ICC analysis provides additional support for this conclusion (Table 2): consistency and

agreement in average performance across perceivers by age group was somewhat lower than in the laboratory condition, though still in the good to excellent range.

Laboratory versus online setting

The final analysis directly compared performance in the laboratory and online setting by adding environment as an additional fixed effect to the mixed-effects model. The results indicated significant effects of environment [$F(1, 476) = 16.11$, $p < .001$] and sentence gate [$F(5, 2380) = 379.92$, $p < .001$] and their interaction [$F(5, 2380) = 7.73$, $p < .001$] on prediction accuracy. There was no effect of speakers’ age group, alone or in combination with other effects. The model R^2 was .38.

The effect of environment indicates that perceivers achieved higher accuracy when they completed the prediction task in the laboratory setting compared to the online setting. The interaction is likely due to the same chance performance across environments on the first two sentence gates and lower accuracy at subsequent gates in the online setting.

A direct comparison of ICC values in the two environments, as a function of the different speaker age groups, indicates overlapping confidence intervals for perceiver agreement and consistency when judging children’s speech, but nonoverlapping intervals when judging adults’ speech (see Fig. 5). More specifically, performance was less reliable on average in an online setting than in a laboratory setting when the gated AV speech that perceivers heard was produced by adults.

The confidence interval also tended to be wider in data collected in the online setting compared to the laboratory setting ($M = .10$ online vs. $M = .04$ in the lab). These results suggest that more perceivers are needed to achieve the same reliability when the gated AV speech method is put online. Table 3 indicates how many more perceivers are needed to achieve an ICC value of .90 with specific confidence intervals ranging from very narrow ($CI = .05$) to fairly broad ($CI = .15$). Note that the calculations are based on the data obtained in the present experiment, including the number of data points per

Fig. 4 Perceiver performance in the forward prediction task is shown as a function of the speakers’ age (i.e., group: adult versus child) and sentence gate at the segmental landmarks indicated, where “æ” indicates the midpoint of the stressed vowel in “packs” and “T” the midpoint of the target vowel (/æ/ or /ou/) in the object noun. Whiskers indicate the 95% confidence intervals, and the dotted line indicates chance performance. The data are from perceivers who completed the task online

Table 2 Agreement and consistency of perceiver performance in the online condition

			Intraclass correlation		95% Confidence intervals		F-test	
					Lower bound	Upper bound	F	p-value
Agreement	Child	0.912			0.865	0.949	12.2	< .001
	Adult	0.871			0.801	0.925	8.58	< .001
Consistency	Child	0.918			0.873	0.952	12.2	< .001
	Adult	0.883			0.820	0.932	8.58	< .001

perceiver (6 gates $\times$ 6 speakers per group = 36 accuracy values). A greater number of data points (e.g., more speakers) per perceiver would change the calculations somewhat, though not dramatically.

Although the specific numbers given in Table 3 apply only to the conditions of the current experiment, it is clear from the table that more perceivers are needed to achieve good to excellent reliability with a high degree of certainty if the gated AV speech method for measuring coarticulation is to be run online. For example, 30 perceivers would have been necessary to achieve an average ICC value of .90 for performance on adult speakers in the online condition with a 95% probability that the true value lie between .85 and .95 (i.e., CI = .10). As it was, we achieved an average ICC value of .88 with 20 perceivers for this group of speakers in the online condition. The 95% confidence interval was also a bit wider (CI = .11). Remarkably few perceivers are needed to achieve very high reliability when the task is administered in the laboratory setting.

Fig. 5 ICC values with 95% confidence intervals for perceiver agreement and consistency in the prediction task when listening to forward-gated AV speech produced by adults and children in the laboratory versus online condition

Discussion

Performance in the forward prediction task indicated that perceivers were able to correctly predict the rhyme of an upcoming noun as early as the onset of a determiner in gated AV speech. Prediction accuracy also improved as the sentence gates approached and then included the target stressed vowel in the noun. Given that acoustic analyses of the determiner vowel also indicated effects of the following vowel on production, we interpret the current results to indicate that the onset of anticipatory coarticulation for an upcoming stressed vowel is at the preceding determiner onset and that the strength of coarticulatory cues to this target vowel increases with decreasing distance from the target.

The present results indicate no effect of speakers' age on perceivers' performance, which suggests that children do not differ from adults in the (normalized) temporal extent over which they prepare for an upcoming rounded vowel—at least given a repetitive elicitation task and a simplified sentence structure and phonological context. However, the acoustic results suggest a clear difference in *how* anticipatory coarticulation extended to preceding segments in child and adult speech. These results are consistent with direct measurement studies of coarticulation that report age-related differences in coarticulatory strategies, but no differences in coarticulatory scope (e.g., Goffman et al., 2008; Rubertus & Noiray, 2018). More importantly, the absence of any significant differences in the consistency and agreement with which perceivers judged child and adult speech strongly suggests that the forward-gated AV speech method can be reliably extended to measure anticipatory coarticulation in children's speech.

Finally, the parallel results across the different task environments suggest that perception-based measurements of coarticulation can be collected online. Nonetheless, more perceivers should be run online to achieve the same levels of measurement reliability as in the laboratory setting.

Experiment 2

Experiment 2 tested whether the gated AV speech method can be extended to measure carryover coarticulation by asking

Table 3 Estimated number of perceivers needed to achieve the listed confidence interval widths given an ICC value of .90 and 36 observations per perceiver (6 speakers per age group $\times$ 6 gates) in the laboratory and online condition

		Confidence interval width										
		.05	.06	.07	.08	.09	.10	.11	.12	.13	.14	.15
Laboratory	Child	19	14	11	8	7	6	5	5	4	4	3
	Adult	12	9	7	5	5	4	4	3	3	3	3
Online	Child	49	35	26	20	16	13	11	10	9	8	7
	Adult	114	80	59	45	36	30	25	21	18	16	14

perceivers to predict a preceding target sound based on backward-gated AV speech. The task used the subset of AV recorded sentences from Kallay and Redford (2018) with the frame, *Maddy*_____ *the gak*, where the verb was either *pack* or *poke*. Thus, the coarticulatory effect under investigation was carryover posturing for the back, rounded mid-vowel /oo/, which was contrasted with maintenance of an overall posture advantageous for the articulation of the low, unrounded front vowel /æ/. Recall that Kallay and Redford (2018) found a long-distance effect of carryover vowel-to-vowel coarticulation on the determiner vowel (schwa) in the larger set of sentences used in that study, but only in children's speech.

The simple subject-verb-*the*-object sentences were gated from the end of the sentence to the midpoint of the stressed vowel in the noun across the unstressed determiner to the midpoint of the target stressed vowel in the verb. Perceivers were to guess the identity of the verb based on the information given. As before, one group of perceivers completed the task in the laboratory, another completed the task online. The prediction was that accuracy would be higher at gates closer to the target vowel in the verb than further away. An effect of age on accuracy was expected based on acoustic and kinematic results that suggest a greater degree of carryover coarticulation in children's speech compared to adults' speech (Kallay & Redford, 2018; see also Rubertus & Noiray, 2020). As before, an effect of condition (laboratory versus online) on accuracy would suggest that perceivers' ability to detect coarticulatory cues varies with the setting. Similarly, an interaction between age and condition on accuracy would suggest a need for caution in interpreting the results with respect to age-related differences in coarticulation. Finally, perceiver agreement and consistency were expected to be equally high across age groups and conditions, provided that backward gating can be used to reliably measure carryover coarticulation. Detailed methods and results follow.

Methods

Participants

Speakers The same 12 speakers who provided sentences for the forward version of the task provided the sentences used in the backward version.

Perceivers Perceivers were 40 new college-aged adults (18–23 years old) recruited from the Linguistics and Psychology human subjects pool. Participants again selected to participate in the current study blind to its purpose and from a wide range of studies to be completed for course credit. Students who had participated in Experiment 1 were restricted from participating in Experiment 2. All of the new participants reported normal hearing and a history of typical speech and language acquisition. Twenty participants completed the task in the laboratory; the other 20 completed the task online. As before, participants were compensated with course credit for their time.

Stimuli

The sentences were elicited as described in Experiment 1. The focus in Experiment 2 was on the sentences with contrasting verbs, and so three good repetitions of the sentences *Maddy pokes the gak* and *Maddy packs the gak* were selected in the manner described previously. This time the sentences were gated to preserve the end of the sentence and eliminate the beginning. The gates therefore determined speech/sentence onset in the video clips that served as stimuli. Speech/sentence onsets were as follows: from the midpoint of the vowel in the verb (Gate 1); from the release of /k/ and onset of /s/ frication in the final cluster of the verb (Gate 2); from the consonantal onset of the determiner (Gate 3); from schwa in the determiner (Gate 4); from /g/ closure in *gak* to the end of the sentence (Gate 5); and from the midpoint of the vowel in *gak* (Gate 6). Figure 6 shows the initial video frame for the six gated stimuli derived from a single child's production of the

Fig. 6 The initial frame from each of six stimuli derived from a single production of the sentence “Maddy packs the goat” by a 5-year-old speaker. Gates were placed at the segmental landmarks shown, beginning with the verb, using acoustic and kinematic cues in the audiovisual recording

sentence *Maddy pokes the gak*. Each sentence repetition was cut in the same way. This resulted in 36 stimuli per speaker (2 sentences $\times$ 3 repetitions $\times$ 6 cuts).

Procedure

The procedure was exactly as in Experiment 1, except that the concept of carryover coarticulation was explained with reference to the example given before. The specific instructions were as follows:

We are not studying you, we are studying the speakers you will be seeing. You are our measurement devices. We are measuring coarticulation, which is an adjustment of how a sound is produced based on the articulation of a preceding or following sound. For example, if you say “stroop” and then say “street” you will feel your lips rounded on the “s” of “stroop” but not on the “s” of “street.” In this study, we are interested in carryover effects; that is, in adjustments to a current sound based on the articulation of a preceding sound. The speakers you will be listening to are saying sentences that have either “pokes” or “packs” as the main verb. The whole sentence is either “Maddy pokes the gak” or “Maddy packs the gak.” You will see clips of them saying these sentences, but the clips are cut so that they start at the midpoint of the verb or later. Your task is to guess whether the preceding verb was “pokes” or “packs.” The shorter the clip, the less information you will have to make your judgment, but do your best and try to decide whether the speaker was saying a sentence with “pokes” or a sentence with “packs.”

Note that the exact same results as those reported here were obtained when the concept of carryover coarticulation was explicitly introduced with reference to liquid devoicing in “please.” Thus, the results appear to be robust to variation in the instruction given.

As in Experiment 1, the response buttons (= “pokes” and “packs”) appeared on the screen with the final frame of the

video. Participants were required to select one as a response before the next stimulus was presented.

Calibration data from the online condition again indicated that video size was controlled in the online condition, and the little variability that did exist was not correlated with performance: Pearson’s $r(20) = -.050$, $p = .83$. Again, no further information was recorded or elicited regarding the at-home conditions in the online setting.

Statistical analyses

Responses were coded as correct or incorrect, and a bias-correct measure of prediction accuracy was calculated for each gate within each speaker and perceiver just as in Experiment 1. Analyses of accuracy and reliability were also as in Experiment 1. Acoustic measurement and analysis of the formant trajectories in the unstressed determiner vowel (schwa) was also performed as in Experiment 1 to provide an acoustic context for the results from the gated AV speech prediction task. These analyses are presented first.

Results

Schwa formant dynamics

Figure 7 presents SS-ANOVA plots of the F1, F2, and F3 trajectories in schwa of the determiner after the verb “packs” versus “pokes” in adult (left) and child (right) speech. The figure suggests that vowel-to-vowel carryover effects were limited to sentences produced by children. The higher F3 values after “pokes” than “packs” likely indicates carryover of the upper pharyngeal constriction for /ou/ during schwa articulation (see Stevens & Keyser, 1989).

Laboratory setting

Turning now to the gated AV speech results, the analysis on perceiver performance in the laboratory setting indicated only an effect of gate on prediction accuracy [$F(5, 1190) = 254.30$, $p < .001$]. The effect of group was not significant alone or in

Fig. 7 SS-ANOVA plots for the schwa in “the” produced by the 6 adult (left) and 6 child (right) speakers as a function of the preceding verb: formants for schwa after “packs” are shown in black; those for schwa after “pokes” are shown in gray. Dotted lines indicate 95% confidence intervals

interaction with gate, despite the age-related acoustic differences noted above. The overall model R^2 was .44.

As expected, perceivers’ ability to correctly predict the verb decreased as the gate increased in distance from the target. Mean percent accuracy by gate, from the target stressed vowel in the main verb through to the fixed stressed vowel in the object noun, was as follows (standard deviation in parentheses): 95.6% (11.8); 65.1% (18.7); 55.0% (19.8); 53.4% (18.2); 52% (17.2); 52.4% (18.4). These results, also evident in Fig. 8, suggest that sufficient information about prior articulation of the mid, back rounded vowel in “poke” (versus “pack”) extended only through the offset of the verb (Gate 2; “k/s” in Fig. 8). Mean comparison tests, corrected for multiple comparison, confirm that accuracy was only significantly higher when the AV clip began either during the offset of the verb or at the midpoint of the target vowel itself (Gate 1; “T” in Fig. 7); there were no significant differences in performance between Gate 3 (“ø” in Fig. 7) and later gates in the sentence. Overall, the perceptual results are at odds with the acoustic results showing carryover effects into the determiner vowel in children’s speech, and so could be interpreted to mean that AV cues to carryover effects from vowel articulation are more limited than AV cues to anticipatory effects.

The ICC analysis revealed excellent reliability with coefficient values of greater than .90 across the different age groups. Table 4 summarizes the results for agreement and consistency. It also presents the result of F -tests for significance by age group for the analysis on average measures.

Online setting

The analysis on perceiver performance in the online setting indicated an effect of group on prediction accuracy [$F(1,$

Fig. 8 Perceiver performance in the backward prediction task is shown as a function of the speakers’ age (i.e., group: adult versus child) and sentence gate at the segmental landmarks indicated, where “T” indicates the midpoint of the target vowel (/æ/ or /ou/) in the verb and “æ” indicates the midpoint of the stressed vowel in the noun “gak.” Whiskers indicate the 95% confidence intervals, and the dotted line indicates chance performance. The data are from perceivers who completed the task in the laboratory

Table 4 Agreement and consistency of perceiver performance in the laboratory condition

		Intraclass correlation	95% Confidence interval		<i>F</i> -test	
			Lower bound	Upper bound	<i>F</i>	<i>p</i> -value
<i>Agreement</i>	<i>Child</i>	0.938	0.905	0.964	17.3	< .001
	<i>Adult</i>	0.945	0.914	0.968	18.4	< .001
<i>Consistency</i>	<i>Child</i>	0.942	0.911	0.966	17.3	< .001
	<i>Adult</i>	0.946	0.916	0.968	18.4	< .001

236) = 4.48, $p = .035$]. The effect of gate was also significant [$F(5, 1185) = 37.59$, $p < .001$] as was the interaction between age group and gate [$F(5, 1185) = 13.70$, $p < .001$]. These results, shown in Fig. 9, are due to poorer perceiver performance in the online backward-gated AV speech task, especially on sentences produced by children. Even when the AV clip included half of the target vowel (Gate 1 or “T” in Fig. 8), mean accuracy in identifying the stressed vowel in children’s speech was only 60.4% ($SD = 37.3\%$). Accuracy was also fairly low at the same gate in adult’s speech ($M = 81.9\%$, $SD = 22.8\%$). Performance at all other gates, including at verb offset (= “k/s”), hovered around chance. Mean comparisons confirmed that only the difference between performance at the first gate (= “T”) and all other gates was significant. The model R^2 was also only .22.

Fig. 9 Perceiver performance in the backward prediction task is shown as a function of the speakers’ age (i.e., group: adult versus child) and sentence gate at the segmental landmarks indicated, where “T” indicates the midpoint of the target vowel (/æ/ or /ou/) in the verb and “æ” indicates the midpoint of the stressed vowel in the noun “gak.” Whiskers indicate the 95% confidence intervals, and the dotted line indicates chance performance. The data are from perceivers who completed the task online

Overall, the results shown in Fig. 9 suggest that substantial noise is introduced into the data when the backward-gated AV speech task is run online. This suggestion is consistent with the findings from the reliability analyses (Table 5), which indicated good—but not excellent—agreement and consistency in perceiver performance.

Laboratory versus online settings

Direct comparison of performance in the laboratory and online settings confirmed the conclusion that speech measurement data collected online using the backward-gated AV speech perception method is significantly noisier than data collected in the laboratory. The overall analyses indicated a significant effect of environment on performance accuracy [$F(1, 474) = 56.99$, $p < .001$]. This effect interacted with the effect of age group and the effect of gate [Environment $\times$ Group, $F(1, 474) = 7.29$, $p = .007$; Environment $\times$ Gate, $F(5, 2375) = 32.57$, $p < .001$; Environment $\times$ Group $\times$ Gate, $F(5, 2375) = 7.44$, $p < .001$]; the simple effect of group was not significant, but the effect of gate was [$F(5, 2375) = 223.99$, $p < .001$]. The overall model R^2 was .38. The simple effect of environment and different interaction effects are evident in Fig. 10, which shows performance accuracy as a function of the three fixed effects in the model.

The comparison of ICC values in the two conditions indicates overlapping confidence intervals for perceiver agreement and consistency across speaker groups (see Fig. 10), but relatively low agreement and consistency in the online condition.

Table 6 shows how many perceivers would be needed to achieve the same ICC value of .90 across conditions and speaker groups given specific confidence intervals. As in Experiment 1, the calculations are based on the data obtained in the present experiment, including the number of data points per perceiver (6 gates $\times$ 6 speakers per group = 36 accuracy values). Unlike Experiment 1, relatively few perceivers are needed to achieve high reliability with a high degree of confidence. This is likely because perceivers were so often performing at chance when trying to identify the target

Table 5 Agreement and consistency of perceiver performance in the laboratory condition

			Intraclass correlation		95% Confidence intervals		F-test	
					Lower bound	Upper bound	F	p-value
Agreement	Child	0.922			0.879	0.954	13.1	< .001
	Adult	0.871			0.802	0.925	8.26	< .001
Consistency	Child	0.819			0.712	0.900	5.53	< .001
	Adult	0.744			0.603	0.850	3.88	< .001

stressed vowel associated with the verb in the sentence. Put another way, the perceivers were all equally confused about the identity of the verb given the backward-gated stimuli.

Discussion

Perceiver performance in the laboratory on the backward prediction task produced results that, at first glance, might suggest that AV cues to carryover coarticulation are more limited than AV cues to anticipatory coarticulation. However, the unexpected finding that perceiver performance varied with speaker age when the task was completed online also ran counter to the acoustic results, which suggested that carryover effects are more pronounced in children's speech compared to adults' speech. Together, the results strongly suggest the need for a validation study to ensure that a perception-derived measure of carryover coarticulation is consistent with more direct measures of coarticulation. Caution is further advised given

the finding that overall performance accuracy is greatly reduced in the online task—to the point where perceivers often failed to identify the target vowel even when the stimuli include the latter half of these vowels. Taken together, the findings from Experiment 2 suggest that the backward version of the task may be too difficult or unnatural to serve as a reliable measure of carryover coarticulation.

General discussion

Psycholinguistic theories of speech production assume a speech plan (e.g., Dell, 1986; Levelt, 1993; Goldrick & Rapp, 2007); phoneticians have long assumed that anticipatory coarticulation indexes this plan (e.g., Nittrouer, Studdert-Kennedy, & McGowan, 1989; Whalen, 1990; Goffman, Smith, & Ho, 2008). If this is correct, then patterns of anticipatory coarticulation can be studied to provide insight into the planning process. If it is not correct, then our theories must address the gap between the underlying psychological and directly measurable aspects of spoken language. The investigation of coarticulatory patterns, particularly long-distance ones, is central to this endeavor. Detailed movement analyses provide detailed and highly individual results, which could hinder the identification of global patterns that are of greater psychological interest. Also, direct measurement methods may adversely impact the naturalness of spoken language, limiting the usefulness of these methods for the psychological study of coarticulation, especially in populations with immature or low inhibitory control. Accordingly, we have offered an alternative, psycholinguistic method for measuring speech movements and sound shaping associated with coarticulation (Redford et al., 2018). The results from the current study indicate that the method is highly reliable when used to measure anticipatory coarticulation regardless of a speaker's age. It is also scalable in that at least the forward-gated version of the AV speech prediction task can be put online. We elaborate on these points below and speculate on what the results imply for implicit speech knowledge and the nature of coarticulation. We also acknowledge the apparent limitations of the method

Fig. 10 ICC values with 95% confidence intervals for perceiver agreement and consistency in the prediction task when listening to backward-gated AV speech produced by adults and children in the laboratory versus online condition

Table 6 Estimated number of perceivers needed to achieve the listed confidence interval widths given an ICC value of .90 and 36 observations per perceiver (6 speakers per age group $\times$ 6 gates) in the laboratory and online condition

		Confidence interval width										
		0.05	0.06	0.07	0.08	0.09	0.10	0.11	0.12	0.13	0.14	0.15
<i>Laboratory</i>	<i>Child</i>	35	25	19	15	12	10	8	7	6	6	5
	<i>Adult</i>	10	7	6	5	4	4	3	3	3	3	2
<i>Online</i>	<i>Child</i>	9	7	5	4	4	3	3	3	3	2	2
	<i>Adult</i>	54	38	28	22	18	15	12	11	9	8	7

for measuring carryover coarticulation and why this is unfortunate for the psychological study of coarticulation. Finally, we briefly discuss the pros and cons of online data collection given the current study results.

Anticipatory coarticulation

The accuracy results from the forward-gated AV speech prediction task (Experiment 1) replicate findings from our prior study (Redford et al., 2018) and easily conform with expectations based on more than a half-century of study on coarticulation, including more recent work: vowel-to-vowel coarticulation spans syllable and word boundaries (e.g., Magen, 1997; Goffman et al., 2008; Grosvald, 2009; Rubertus & Noiray, 2018); it may however be “blocked” by stressed syllables (see, e.g., Fowler, 1981). The accuracy results also suggest minimal to no age-related difference in coarticulatory scope given a simple elicitation task. This finding is consistent with recent findings from kinematic studies of coarticulation in children’s speech (e.g., Goffman, Smith, & Ho, 2008; Rubertus & Noiray, 2018).

The sensible accuracy results were coupled with very excellent reliability results. Perceiver agreement and consistency were also invariant as a function of the speaker’s age. This latter finding is important for the methodological reason that motivated the current study, namely, the need to develop a noninvasive and nonrestrictive method for studying long-distance coarticulation in children’s speech. The results are also interesting for what they suggest about our implicit knowledge of speech. For example, although children’s speech is different than adults’ speech due to the protracted development of speech motor control, college-aged participants reliably identified articulatory pre-posturing for an upcoming vowel at the beginning of the determiner in both 5-year-old and adult speech. This finding is all the more impressive under the reasonable assumption that most of our college-age participants had minimal daily experience with young children’s speech. Overall, the results suggest that perceivers weigh a variety of cues to articulation, including both audio and visual sensory information, to make judgments about

upcoming speech sounds, perhaps because they are implicitly aware of articulatory synergies, and that, whatever the internal model of articulation is, it is robust enough to make systematic forward predictions about speech given highly variable input.

Carryover coarticulation

In contrast to the results from the forward version of the task, the overall accuracy results from the backward-gated AV speech prediction task (Experiment 2) should probably not be interpreted as a measure of coarticulation. Whereas the schwa formant dynamics indicate stronger carryover effects in children’s speech compared to adults’ speech, consistent with kinematic study results (e.g., Rubertus & Noiray, 2020), no effect of age group was observed in the results from data collected in the laboratory setting. In addition, the results from data collected online showed that perceiver performance varied systematically with age, albeit in the opposite direction from what one might expect given the acoustic results. Moreover, perceivers who completed the backward task online performed poorly in the prediction task even when a portion of the target vowel was provided. In these ways, the online results did not replicate the laboratory results, which leads us to question the validity of the assumption that perceiver performance in the backward-gated AV speech task is based on coarticulatory cues.

Unlike anticipatory coarticulation, which is linked at some level in the production process to planning, carryover coarticulation is typically explained as the result of inertial effects on the articulators (see Whalen, 1990; Magen, 1997:189). If we are correct to assume that listeners bring their knowledge as speakers to bear in the forward- and backward-gated AV speech prediction tasks studied here, then it could be that perceiver performance is poorer in the backward version of the task compared to the forward version because carryover effects are not attached to production units that the perceiver aims to extract from the speech stream. Moreover, if carryover effects are merely inertial effects, then they are not really of psychological interest, at least when the goal of study is to better understand the production planning process.

On the other hand, some studies indicate language-specific differences in the strength and scope of carryover effects (e.g., Beddor, Harnsberger, & Lindermann, 2002; Mok, 2010). Such findings suggest that other forces besides articulatory inertia influence the persistence of an articulatory posture associated with a sound that has already been achieved. Although these forces may still fall shy of direct control over carryover coarticulation—for example, the language-specific effects may be an epiphenomenon of prosodic differences (Mok, 2010)—the importance of inhibition to sequencing operations argue for further investigation of carryover effects from a psychological perspective. For the moment, that investigation should probably be based on acoustic measurement, which is less invasive than direct measurement. Again, the backward version of the psycholinguistic method presented here requires further validation before being used to measure carryover coarticulation.

Task setting

The study included a comparison of data collected in the laboratory compared to online to assess whether the psycholinguistic method we have developed generalizes across settings, and to determine whether the method can be streamlined to investigate coarticulation in larger samples of speakers from different age groups. After all, online data collection is much less time-consuming than in-laboratory data collection. The comparison had the added benefit of providing an internal quasi-replication of Experiments 1 and 2. The results from the forward version of the task strengthen the confidence with which we assert that the forward-gated AV speech task provides a robust and reliable measure of long-distance anticipatory coarticulation, whereas the absence of replication in the backward version of the task in the online condition contributes to the conclusion that the task cannot be used to measure long-distance carryover coarticulation. Yet, even in the forward version of the task, we found that overall magnitude of the main effect of gate on perceiver performance was reduced. This is consistent with some prior work showing poorer online than in-laboratory performance in psychological experiments (e.g. Miller, Crowe, Weiss, Maples-Keller, & Lynam, 2017). It is inconsistent with other prior work that indicates better attention and performance online than in the laboratory (Hauser & Schwarz, 2016). The conflicting results argue for further comparison of laboratory and online methods of data collection. Certainly, the current results support putting the forward-gated version of the task online, but also argue for increasing the number of perceivers used per speaker to measure anticipatory coarticulation. All possible information regarding at-home setup should also be recorded and/or elicited from participants in follow-up online research. This

information could provide insight into the source of performance variability in the online setting, and would at least allow for statistical control over this source.

Conclusion

The purpose of the current study was to assess the reliability and generalizability of a psycholinguistic method for measuring coarticulation. The method leverages AV speech perception to provide more information than acoustic measurement about the scope and strength of coarticulation in running speech. But the use of perception to measure production begs the question of confounding influences on perceiver performance. Here, we tested for several such influences by examining the effects of the speaker's age, task setting, and naturalness on perceiver performance. The results from Experiment 1 indicated good-to-excellent inter-perceiver reliability in the forward-gated version of the task in the laboratory and online settings regardless of the speakers' age. The results from Experiment 2 indicated an effect of task naturalness on perceiver performance. In particular, perceiver performance on most gates hovered around chance in the backward-gated version of the task. This result might suggest that AV cues to carryover coarticulation are less robust than those to anticipatory coarticulation, but it could also indicate that the backward prediction task was difficult because it is much less natural than the forward prediction task. Further, the results from both experiments indicate a decrement in perceiver performance when the prediction tasks are run online. But, whereas online performance in the forward-gated version of the task replicated the results from the laboratory setting, suggesting that the minor performance decrement can be overcome by adding perceivers, online performance in the backward-gated version of the task resulted in effects not seen in results from the laboratory setting. Taken together, the findings strongly support the use of the gated AV speech method for measuring the onset and strength of anticipatory coarticulation in child and adult speech, but also suggest that the method not be extended to measure carryover coarticulation absent its validation with direct measurement.

Acknowledgements This research was wholly supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development (NICHD) under grant R01HD087452 (PI: Redford). The content is solely the authors' responsibility and does not necessarily reflect the views of NICHD.

Open Practices Statement The data and materials of this study are available on reasonable request from the corresponding author [MAR]. The data and materials are not publicly available due to privacy restrictions. The experiments were not preregistered.

References

- Arditte, K. A., Çek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment*, 28, 684–691.
- Bates, D. Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Baum, S. R., & McFarland, D. H. (2000). Individual differences in speech adaptation to an artificial palate. *Journal of the Acoustical Society of America*, 107(6), 3572–3575.
- Beddor, P., Harnsberger, J., and Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates, *Journal of Phonetics*, 30, 591–627.
- Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493), 1001–1010.
- Davidson, L. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance. *The Journal of the Acoustical Society of America*, 120(1), 407–415.
- Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3), 283.
- Dell, G. S., Schwartz, M. F., Martin, N. M., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, 104(4), 801–838.
- Derrick, D., & Gick, B. (2011). Individual variation in English flaps and taps: A case of categorical phonetics. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 56(3), 307–319.
- Dunn, L., & Dunn, L. (2007). *Peabody Picture Vocabulary Test* (4th). Circle Pines, MN: American Guidance Service.
- Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44, 491–505.
- Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech and Hearing Research*, 24(1), 127–139.
- Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). Package ‘irr’: Various Coefficients of Interrater Reliability and Agreement. <https://CRAN.Rproject.org/package=irr>.
- Goffman, L., Smith, A., Heisler, L., Ho, M. (2008). The breadth of coarticulatory units in children and adults. *Journal of Speech, Language, and Hearing Research*, 51, 1424–37.
- Goldman, R. & Fristoe, M. (1999) *Goldman-Fristoe Test of Articulation 2*. Circle Pines, MN: American Guidance Service Inc.
- Goldrick, M., & Rapp, B. (2007). Lexical and post-lexical phonological representations in spoken production. *Cognition*, 102, 219–260.
- Goodman, J. K., Cryder, C. E. & Cheema, A. (2013). Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioural Decision Making*, 26, 213–224.
- Grosvald, M. (2009). Interspeaker variation in the extent and perception of long-distance vowel-to-vowel coarticulation. *Journal of Phonetics* 37(2), 173–188.
- Hauser, D. J. & Schwarz, N. (2016). Attentive turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48, 400–407.
- Jordan, T., Sergeant, P., Martin, C., Thomas, S., & Thow, E. (1997). Effects of horizontal viewing angle on visual and audiovisual speech perception. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation* (Vol. 2, pp. 1626–1631). IEEE.
- Jordan, T. R., & Bevan, K. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 23(2), 388–403.
- Jordan, T. R., & Sergeant, P. C. (1998). Effects of facial image size on visual and audio-visual speech recognition. In Campbell R., Dodd, B., & Burnham, D. (eds.), *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*, pp. 155–176. Hove, England: Psychology Press.
- Jordan, T.R., & Thomas, S. M. (2001). Effects of horizontal viewing angle on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 27(6), 1386–1403. <https://doi.org/10.1037/0096-1523.27.6.1386>.
- Kallay, J., & Redford, M.A. (2018). Coarticulatory effects on “the” production in child and adult speech. *Proceedings of the 9th International Conference on Speech Prosody 2018*, 1004–1007.
- Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 155–163.
- Lenth, R. (2018). *emmeans: Estimated Marginal Means, aka Least-Squares Means* [R package]. Version 1.3.0.
- Levelt, W. J. (1993). *Speaking: From Intention to Articulation*. Boston, MA: MIT press.
- Magen, H. S. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics* 25, 187–205.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748.
- Miller, J. D., Crowe, M., Weiss, B., Maples-Keller, J. L., & Lynam, D. R. (2017). Using online, crowdsourcing platforms for data collection in personality disorder research: the example of Amazon’s Mechanical Turk. *Personality Disorders: Theory, Research, and Treatment* 8(1), 26–34.
- Mok, P. K. P. (2010). Language-specific realizations of syllable structure and vowel-to-vowel coarticulation. *Journal of the Acoustical Society of America*, 128(3), 1346–1356.
- Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech, Language, and Hearing Research*, 32(1), 120–132.
- Noiray, A., Abakarova, D., Rubertus, E., Krüger, S., & Tiede, M. (2018). How do children organize their speech in the first years of life? Insight from ultrasound imaging. *Journal of Speech, Language, and Hearing Research*, 61, 1355–1368.
- Noiray, A., Ménard, L., & Iskarous, K. (2013). The development of motor synergies in children: Ultrasound and acoustic measurements. *Journal of the Acoustical Society of America*, 133, 444–452.
- Pickering, M. J. & Garrod, S. (2007) Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11, 105–110.
- R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66(336), 846–850.
- Redford, M. A., Kallay, J. E., Bogdanov, S. V., & Vatikiotis-Bateson, E. (2018). Leveraging audiovisual speech perception to measure anticipatory coarticulation. *Journal of the Acoustical Society of America*, 144, 2447–2461.
- Rubertus, E., & Noiray, A. (2018). On the development of gestural organization: A cross-sectional study of vowel-to-vowel anticipatory coarticulation. *PloS One*, 13(9), e0203562.
- Rubertus, E., & Noiray, A. (2020). Vocalic activation width decreases across childhood: Evidence from carryover coarticulation. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 11(1), 7. DOI: <https://doi.org/10.5334/labphon.228>
- Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, 71(1), 145–163.

- Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, 1, 213–220.
- Smith, A., & Zelaznik, H. (2004). The development of functional synergies for speech motor coordination in childhood and adolescence. *Developmental Psychobiology*, 45, 22–33.
- Stevens, K. & Keyser, S. J. (1989). Primary features and their enhancement in consonants. *Language*, 65, 81–106.
- Warren, R. M. (1970) Perceptual restoration of missing speech sounds. *Science* 167, 392–93.
- Westbury, J. R., Hashi, M., & Lindstrom, M. J. (1998). Differences among speakers in lingual articulation for American English /r/. *Speech Communication*, 26, 203–226.
- Whalen, D. (1990). Coarticulation is largely planned. *Journal of Phonetics*, 18, 3–35.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag: New York.
- Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for Estimating Repeatability. *Methods in Ecology and Evolution*, 3(1), 129–137.
- Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1–2), 23–43.
- Zharkova, N., Hewlett, N., & Hardcastle, W. J. (2008). An ultrasound study of lingual coarticulation in children and adults. *Proceedings of the International Speech Production Seminar*, 161–164.
- Zharkova, N., Hewlett, N., & Hardcastle, W. J. (2011). Coarticulation as an indicator of speech motor control development in children: an ultrasound study. *Motor Control*, 15, 118–140.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.