Leveraging audiovisual speech perception to measure anticipatory coarticulation

Melissa A. Redford, Jeffrey E. Kallay, Sergei V. Bogdanov, and Eric Vatikiotis-Bateson

Citation: The Journal of the Acoustical Society of America **144**, 2447 (2018); doi: 10.1121/1.5064783
View online: https://doi.org/10.1121/1.5064783
View Table of Contents: https://asa.scitation.org/toc/jas/144/4
Published by the Acoustical Society of America

Leveraging audiovisual speech perception to measure anticipatory coarticulation

Melissa A. Redford,[1,a)] Jeffrey E. Kallay,[1] Sergei V. Bogdanov,[1] and Eric Vatikiotis-Bateson[2]
[1]*Department of Linguistics, University of Oregon, Eugene, Oregon 97403, USA*
[2]*Department of Linguistics, University of British Columbia, Vancouver, British Columbia, Canada*

(Received 27 December 2017; revised 27 September 2018; accepted 7 October 2018; published online 29 October 2018)

A noninvasive method for accurately measuring anticipatory coarticulation at experimentally defined temporal locations is introduced. The method leverages work in audiovisual (AV) speech perception to provide a synthetic and robust measure that can be used to inform psycholinguistic theory. In this validation study, speakers were audio-video recorded while producing simple subject-verb-object sentences with contrasting object noun rhymes. Coarticulatory resistance of target noun onsets was manipulated as was metrical context for the determiner that modified the noun. Individual sentences were then gated from the verb to sentence end at segmental landmarks. These stimuli were presented to perceivers who were tasked with guessing the sentence-final rhyme. An audio-only condition was included to estimate the contribution of visual information to perceivers' performance. Findings were that perceivers accurately identified rhymes earlier in the AV condition than in the audio-only condition (i.e., at determiner onset vs determiner vowel). Effects of coarticulatory resistance and metrical context were similar across conditions and consistent with previous work on coarticulation. These findings were further validated with acoustic measurement of the determiner vowel and a cumulative video-based measure of perioral movement. Overall, gated AV speech perception can be used to test specific hypotheses regarding coarticulatory scope and strength in running speech.
https://doi.org/10.1121/1.5064783

[LK] Pages: 2447–2461

I. INTRODUCTION

State-of-the-art methods for measuring speech kinematics include point-tracking systems, such as Optotrak (Northern Digital Inc., Waterloo, Ontario, Canada) or electromagnetic articulography (EMA), and imaging techniques, such as ultrasound and real-time magnetic resonance imaging (MRI). The point-tracking methods provide rich information about inter-articulator coordination. Ultrasound provides detailed information about the movements of specific speech articulators (i.e., the tongue). Real-time structural MRI provides information on both. These methods are required to answer specific questions about articulatory biomechanics and physiology; they can also be used to address questions pertaining to the development of speech motor control (e.g., Goffman and Smith, 1999; Green *et al.*, 2000; Noiray *et al.*, 2013; Zharkova *et al.*, 2011). But direct methods for precisely measuring speech kinematics also have well-known drawbacks. Apart from their expense, they encumber or constrain the speaker in some way. Point-tracking systems require the adhesion of wired sensors or reflective markers to a speaker's face and speech articulators; ultrasound requires head stabilization and a probe to be held under the speaker's chin; MRI requires that the speaker be confined, supine, and that his/her head be immobilized. These drawbacks adversely affect the naturalness of speech production (Stone, 2010), and so may perturb the psychological processes that underlie it. This could be especially true in children and other speakers with poor inhibitory control who are more likely than healthy, neurotypical adults to be distracted by the physical context in which they are speaking. Moreover, our view is that a focus on the decontextualized movements of individual articulators limits inference about the underlying psychology of speech production, including inferences about the speech plan (or planning). This view is simply the complement of the previously stated view that a focus on articulatory subsystems is critical for understanding biomechanical constraints and speech physiology. The issue is one of granularity: if the research questions are psychological, then it is better to abstract away from the motor forces that influence articulatory-specific kinematics in order to focus on the initiation and extent of speech action in relation to linguistic units or processes. Since our research interest is in the psychology of speech production, and especially in its development, we propose a noninvasive method for the synthetic measurement of speech kinematics based on audiovisual (AV) speech perception. The method is presented here by way of a validation study with adult speakers.

A. Measuring production with perception

Our specific interest is in the measurement of long-distance anticipatory coarticulation, a behavior that has long been hypothesized to reflect information about the speech plan (e.g., Jordan, 1997; Whalen, 1990; *inter alia*). In so far as direct measurement methods may disturb the naturalness

a)Electronic mail: redford@uoregon.edu

of speech production and synthetic measures of speech movement are preferable to articulatory-specific ones for answering questions about the psychology of speech production, we have concluded that the best approach for measuring long-distance coarticulation is to use perceptual judgments. Katz *et al.* (1991) seem to have arrived at a similar conclusion well before us. Their study of coarticulation in children's and adults' speech was largely based on perceptual judgments. But whereas Katz and colleagues limited the information they provided to perceivers by cropping video stills so that only static information about mouth shape remained, the present study seeks to validate a method that leverages perceivers' ability to synthesize perioral and other relevant facial and head movements in a running speech context to identify the scope and strength of anticipatory behavior.

AV speech perception is the ideal tool for making production measures that index representation precisely because it abstracts away from the production variance introduced by speech motor control and individual differences in speech physiology. For example, the perceiver's synthesis of visual and auditory information overcomes the problem of motor equivalence, which renders individual articulators unreliable indices of more abstract representations. In particular, speaker adaption to articulator-specific perturbations (e.g., Kelso *et al.*, 1984; Savariaux *et al.*, 1995; Tremblay *et al.*, 2003) amply demonstrates that individual motor goals can be realized differently across contexts. Similarly, individual differences in the articulation of target speech sounds demonstrate the potential problems of inferring representation from direct measurement of the articulators alone. By contrast, perceivers typically recover a speaker's intended goal no matter the individual differences in the details of articulation and ensuing acoustics. This well-known phenomenon, known as the *lack of invariance problem*, is central to the theoretical study of speech perception and production (see, e.g., Perkell and Klatt, 1986). In addition, perceivers are sensitive to the fact that speech involves more than the coordination of the glottis, lips, tongue, and velum: it also involves correlated movements of the face (Munhall and Vatikiotis-Bateson, 1998; Yehia *et al.*, 2002), rhythmic movements of the head (Hadar *et al.*, 1983; Munhall *et al.*, 2004; Thomas and Jordan, 2004), and, of course, large timescale movements associated with respiration (Fuchs *et al.*, 2013; Huber, 2008; Winkworth *et al.*, 1995). These movements may be as important to the recovery of prosodic structure as the coordinated movements of the oral articulators are to the recovery of segmental articulation.

Given our interest in coarticulation, it is also important to note that facial movements are a surprisingly good proxy for the coordinated movements of oral speech articulators and therefore of segmental articulation. In particular, movements of the lip, tongue, and jaw deform the soft tissues of the face in the perioral region so completely that its measurement predicts speech acoustics with the same level of accuracy as measurement of the speech articulators themselves (Yehia *et al.*, 1998, p. 41). This point deserves repetition: coordinated movements that involve the tongue—a hidden articulator—can be recovered from the deformation of soft tissues in the perioral region of the face. The correlation between facial movement and speech acoustics is no doubt why study after study in AV speech perception find a significant processing advantage for visually presented speech in noise compared to audio-only presentation (e.g., MacLeod and Summerfield, 1987; Ross *et al.*, 2006; Summerfield, 1992). In fact, the integration of visual and auditory information is so fundamental to speech processing that significant confusion results when visible speech movements and speech acoustics do not match (e.g., McGurk and MacDonald, 1976).

In the present study, AV speech perception is used to identify the onset of movements associated with long-distance anticipatory coarticulation to test specific hypotheses regarding the scope of a planned production unit. More specifically, degree of coarticulation at pre-determined locations is assessed based on perceivers' ability to predict an upcoming vowel/rhyme in gated AV speech where the gates correspond to the locations. Of course, if this method is to be at all useful for measuring long-distance coarticulation, it must also return results that are expected based on well-established findings in the literature on anticipatory vowel coarticulation. The current study focuses on three such findings in order to validate the method.

B. Anticipatory vowel coarticulation

One well-established finding in the adult literature on anticipatory vowel coarticulation is the effect of a subsequent vowel on the articulation of a current vowel across a syllable boundary. The study by Öhman (1966) on vowel formants in vowel-consonant-vowel (VCV) sequences is the classic reference for this effect, but the result has been replicated many times. Subsequent work has shown that vowel-to-vowel coarticulation is hardly limited to the VCV context. For example, Magen (1997) found that vowel-to-vowel influences can extend across an intervening unstressed syllable in trisyllabic nonwords (i.e., CVCəCVC, where C = consonant and V = full vowel); specifically, all four speakers in her study showed effects of the second full vowel on production of the first (i.e., a significant effect of *V*2 on *F*2 of *V*1).

Magen's (1997) study was undertaken to test the psycholinguistic hypothesis that metrical structure defines planned production units (see, e.g., Roelofs and Meyer, 1998), which in turn define coarticulatory domains under the assumption of a plan-to-coarticulation relation. In English, an unstressed syllable is footed with a preceding stressed syllable (Hayes, 1982). This means that in Magen's trisyllabic nonwords a foot boundary intervened between the weak syllable and the final syllable. Thus, her results showed that a metrical foot boundary does not impede coarticulatory effects, at least when the boundary is located within a nonce word. Grosvald (2009) extended Magen's investigation to real word sentences. His focus was on scope rather than on metrical structure, but the results can be interpreted to suggest that coarticulation is limited by a metrical foot boundary when this coincides with a real word boundary. This prediction is tested in the current study.

Grosvald (2009) recorded 20 speakers repeatedly saying 1 of 2 sentences: "It's fun to look up at a key" and "It's fun to look up at a car." The analysis investigated effects of the vowel in the target nouns "key" and "car" on the reduced vowels of the verb particle "up," the preposition "at," and the indefinite determiner "a." Although metrical structure was not explicitly manipulated, metrical theory predicts that "up" is footed with the preceding strong syllable "look" but that "at" and "a" are extrametrical. In other words, Grosvald's sentences included a metrical foot boundary after the particle "up," but not necessarily after the preposition "at" and certainly not after the indefinite determiner "a." The results were that all 20 speakers in Grosvald's study showed an effect of the vowel from the target noun on the production of the indefinite determiner; 15 of 20 speakers showed an effect of the vowel from the target noun on "at"; only 2 of 20 speakers showed an effect on "up." This pattern of results suggests an effect of metrical foot boundaries, albeit one confounded with distance—the factor of interest to Grosvald. The present study includes a manipulation of metrical boundaries at word boundaries with distance from the target held constant to directly test the prediction that a metrical foot boundary limits the scope of coarticulation when this coincides with a real word boundary.

A second key finding on anticipatory vowel coarticulation is that it influences the production of consonants as well as vowels. This effect is easily intuited by attending to the shape of one's lips while producing the sound /s/ in the words "sack" and "soup"—they are relaxed for "sack" and constricted for "soup." Studies of adult speech kinematics have shown that this effect can extend across many consonants (e.g., "strap" vs "stroop"), including those that cross a word boundary (Kozhevnikov and Chistovich, 1965; Daniloff and Moll, 1968; Bell-Berti and Harris, 1979; *inter alia*). Thus, if one is to measure coarticulation using perception, visual cues will be important since these can be used to detect anticipatory effects at consonantal word onsets or offsets that are not recoverable when consonants are poorly audible (e.g., interdental fricatives) or not at all audible (e.g., stop consonants). Relatedly, work in AV speech perception has shown that visual cues to speech articulation are available earlier in the speech stream than acoustic cues (Cathiard *et al.*, 1996; Munhall and Tohkura, 1998; Moradi *et al.*, 2013). For example, Munhall and Tohkura (1998), citing Smeele (1994), report that "visual information (is) useful up to 150 milliseconds prior to the acoustic onset for the syllable (p. 532)." The present study compares perceiver performance in a gated AV speech task and a gated audio-only speech task to test both for effects of anticipatory behavior on consonantal articulation and the prediction that perceivers will detect these in AV speech, but not necessarily in audio-only speech.

A third key finding in the literature on anticipatory vowel coarticulation is that some consonants are more resistant to coarticulatory effects than others (Bladon and Al-Bamerni, 1976; Bladon and Nolan, 1977). Vowel-to-vowel coarticulation is also affected by coarticulatory resistance (Recasens, 1984; Recasens *et al.*, 1997). For example, Recasens (1984) showed that consonants whose articulation is dependent on the tongue dorsum (e.g., dorsopalatals) will block vowel-to-vowel effects more strongly than consonants whose articulation is primarily dependent on the tongue tip (e.g., /n/). The idea is that coarticulatory resistance emerges when different commands are issued to the same articulator: both dorsopalatal consonants and vowels are articulated using the tongue dorsum. Given this understanding, it is possible to extend the concept of resistance to segments that interfere to a greater or lesser degree with typical patterns of inter-articulatory coordination. For example, anticipatory jaw lowering for an upcoming low vowel is likely to be reduced in the presence of an alveolar consonant, where tongue tip raising is accompanied by a fronted tongue dorsum and a synergistic, high jaw position (see, e.g., Keating *et al.*, 1994); it is likely to be less impacted by the presence of a velar consonant because English allows tongue dorsum contact with the palate to vary substantially with the subsequent vowel, which in turn allows greater degrees of freedom for synergistic movement of the jaw. Here, we test for effects of inter- and intra-articulatory constraints on anticipatory coarticulation by varying the consonantal onset to the target noun. The prediction is that perceivers will more accurately predict the upcoming rhyme when the noun onset has low or no coarticulatory resistance (e.g., /h/) than when it has high coarticulatory resistance (e.g., /s/). Similarly, when the noun onset has an intermediate degree of coarticulatory resistance (e.g., /g/) accuracy should be lower than when there is no resistance, but higher than when there is strong resistance.

C. Current study

The aim of the current study was to validate a noninvasive method for the measurement of anticipatory coarticulation based on AV speech perception. Speakers were audio-video recorded while producing simple sentences with determiner phrases in sentence object position. The determined noun was monosyllabic and had either a rounded or unrounded rhyme (e.g., "soup" vs "sack"). This was the coarticulatory target. The verb was either a plain monosyllabic verb ("pack") or a disyllabic phrasal verb ("pack up"). The metrical boundary thus occurred *after* the determiner in the plain verb sentences and *before* the determiner in the phrasal verb sentences. Noun onset was varied to manipulate coarticulatory resistance. The scope and strength of coarticulation was then assessed at predetermined temporal locations (=gates). These locations were defined with respect to linguistic material to test the prediction that coarticulation is limited by a metrical foot boundary when this coincides with a real word boundary. The critical gates were the onset of the determiner, the determiner vowel, and the onset of the object noun. AV and audio-only versions of the gated stimuli were presented to perceivers whose task was to guess, based on the fragments presented, whether the sentence-final word rhymed with "oop" or "ack." The prediction was for above chance performance at the determiner onset in the phrasal verb sentences, but not in the plain verb sentences where above chance performance was not expected until the onset of the noun. In addition to the metrical boundary prediction, we expected that (a)

perceivers in the AV condition would accurately predict the final rhyme at earlier gates than perceivers in the audio-only condition, and (b) that prediction accuracy (=strength) would vary with the coarticulatory resistance of object noun onsets. These expectations follow directly from the literature on anticipatory vowel coarticulation.

In addition to replicating well-known findings, the perceptual results should parallel those that can be obtained from directly measuring speech acoustics and movement in the same video fragments. Accordingly, the acoustics of the determiner vowel and speech movement across the sentences were also analyzed. Since AV speech perception provides a highly synthetic measure of speech movement, we chose an equally synthetic, video-based direct measure: optical flow analysis (Barbosa *et al.*, 2008). This measure also allowed us to define coarticulation by analogy to the perception task as the difference in cumulative movement profiles for comparable "oop" and "ack" sentences. The complete method details for the study are provided next.

II. METHOD

A. Participants

Five college-aged females were recruited by word-of-mouth from linguistics classes at the University of Oregon to serve as speakers. All spoke the West Coast variety of American English that is standard in Oregon as their native language, and also reported normal hearing. Speakers were financially compensated for their time. In addition, 85 college-aged students (47 female) were recruited from the University of Oregon Psychology and Linguistics Human Subjects Pool to serve as perceivers. All were native speakers of English and compensated with course credit for their time.

B. Materials

1. Speech stimuli

The stimuli were subject-verb-object sentences with target determiner phrases in sentence object position. All sentences began with "Maddy" and all ended in a monosyllabic noun with either a target high, back, rounded vowel (=/u/) or low, front, unrounded vowel (=/æ/). The /u/ nouns had the voiceless bilabial stop /p/ in coda position and so an "oop" rhyme. The /æ/ nouns had either the voiceless alveolar or velar stop /t, k/ in coda position and so an "at" or "ack" rhyme. Noun onsets were varied to manipulate the degree of coarticulatory resistance from none (=hoop/hat) to some (=goop/gak) to strong (=soup/sack). Recall that resistance was defined with reference to the articulatory constraints provided by inter- as well as intra-articulatory coordination and not simply with respect to tongue dorsum constraints (cf. Recasens, 1984). All nouns were introduced with the definite article "the," which was the main focus of anticipatory effects in the design and analyses. Metrical context for "the" was varied using a phrasal verb ("pack up") or plain verb ("pack"). Assuming that "pack up" forms its own trochee, "the" should be prosodified with the noun in the phrasal verb context and with the preceding verb in the plain verb context. Thus, the boundary between "the" and target noun was assumed to be prosodically weaker in the phrasal verb context than in the plain verb context.

2. Elicitation procedure

Sentences were blocked by verb. Target nouns were randomized within the block. Speakers sat in front of a blue screen and facing a Panasonic HC-V770 audio-video camcorder (Panasonic Corp., Kadoma, Osaka, Japan). A large window behind the camcorder provided natural light, which was supplemented by overhead fluorescent lighting. Video recording speed was the standard 30 frames per second (fps). Speakers wore a Shure ULX1-M1 lavaliere wireless microphone (Niles, IL) attached to a headband; the Shure ULXS4 receiver was plugged directly into the camcorder to ensure quality audio that was seamlessly aligned with the video. The experimenter was also female and a native speaker of the same West Coast variety of American-English as the speakers. She stood off to the side of the camcorder and cued production by saying the target sentence, counting slowly to three, at which point the speaker was to produce the target sentence. This method was used to ensure that the speaker looked directly at the camera when speaking rather than down at her lap or to the side to read a stimulus sentence. The delayed repetition was intended to minimize the likelihood that the speaker was modeling the experimenter. Each target noun was elicited once per block. Seven repetitions of each sentence were elicited.

3. Gating procedure

Five good productions of each target sentence were identified, starting with the second repetition. Good productions were those where the speaker produced the sentence under a single intonation contour, without pausing, and without smiling, laughing, or coughing, etc., during speaking.

Once identified, target sentences were extracted from the audio-video recordings and displayed as a sequence of picture frames time-aligned with an oscillogram of the audio in Final Cut Pro Version 7, which is no longer supported by Apple, Inc. (Cupertino, CA).[1] Gates were then identified by toggling from frame-to-frame using acoustic and kinematic landmarks associated with segments of interest. Specifically, acoustic correlates of a segment were identified and the video cut to include a frame associated with the target production of that segment (see Fig. 1). If the duration of the acoustic correlate in focus spanned more than one frame (e.g., stop closure), then a frame was chosen based on the apex of movement (e.g., maximal lip compression), which was defined perceptually by toggling back and forth through the frames in question. Using these procedures, the phrasal verb sentences were gated as follows: the first gate was at the bilabial closure for "up," during maximal compression of the lips (gate a); the second gate was during consonantal closure for "the" (gate b); the third was at the midpoint of the schwa vowel for "the" (gate c); the fourth was during the initial consonant of the target noun, that is, during the /s/ frication, /g/ stop closure, and /h/ aspiration (gate d); and the fifth was at the midpoint of the target noun vowel (gate e). The plain verb sentences had an additional cut, but were

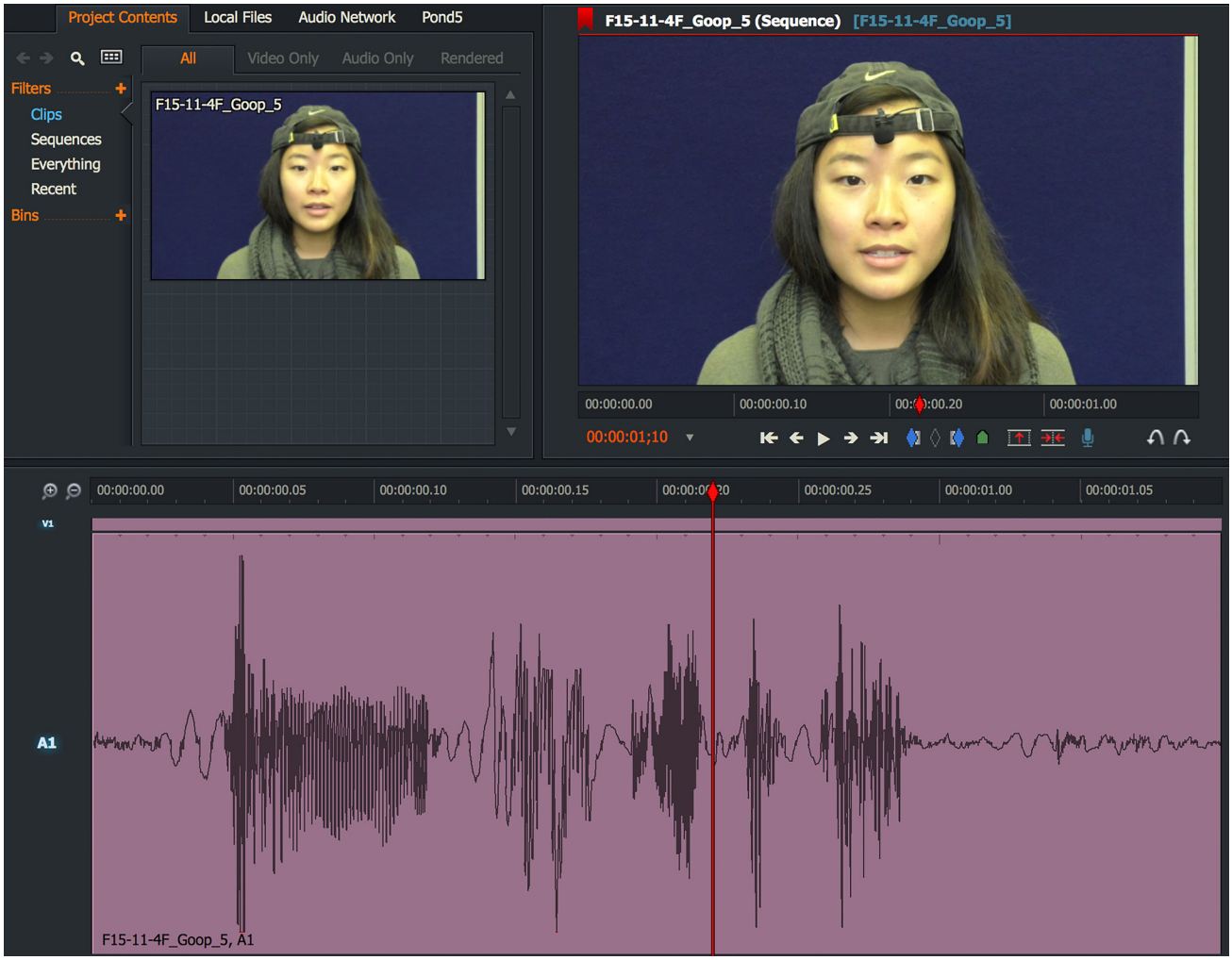

FIG. 1. (Color online) Audio-video recorded sentences were gated at locations that corresponded to specific segmental articulations, which were determined primarily based on acoustic correlates. Here, the gate corresponding to the determiner onset is shown for a production of "Maddy packs the goop."

otherwise similarly gated: the first gate was at the midpoint of the vowel in "pack" (gate a); the second was between the consonantal offset to the verb and the third person marking (gate b); the third was during consonantal closure for "the" (gate c); the fourth was at the midpoint of the schwa vowel for "the" (gate d); the fifth was during the initial consonant of the target noun (gate e); and the sixth was at the midpoint of the target noun vowel (gate f). Since recording speed was 30 fps, the cuts were as close as possible to the midpoint of the segment of interest within a 33-ms interval. Figure 1 illustrates the acoustic-to-frame alignment for determiner onset. Table I summarizes the gate locations for the different sentence types. Figure 2 illustrates the result of the gating procedure by showing the final frame of six clips created by repeatedly cutting an audio-video recording of one speaker producing the sentence "Maddy packs the goop." Table II gives the intervals between gates in number of frames as a function of sentence type for each of the five speakers. Absolute duration can be derived for the intervals by multiplying the given numbers by 33 ms.

Each of the five repetitions of every target sentence was cut in the same way. Since a clip of the whole sentence was also always included in the presentations (labeled "f" for phrasal verb sentences and "g" for plain verb sentences), there were 180 stimuli per speaker for phrasal verb sentences (6 sentences × 5 repetitions × 6 clips) and 210 stimuli per speaker for the plain verb sentences (6 sentences × 5 repetitions × 7 clips). The audio-only condition was created by simply extracting the sound files from each of the audio-video clips created during the gating procedure.

TABLE I. Gate locations for the different sentence types.

	Gate location						
Sentence type	a	b	c	d	e	f	g
Phrasal verb	"up" C	"the" C	"the" V	Noun C	Noun V	Full sentence	
Plain verb	"pack" V	"pack" C	"the" C	"the" V	Noun C	Noun V	Full sentence

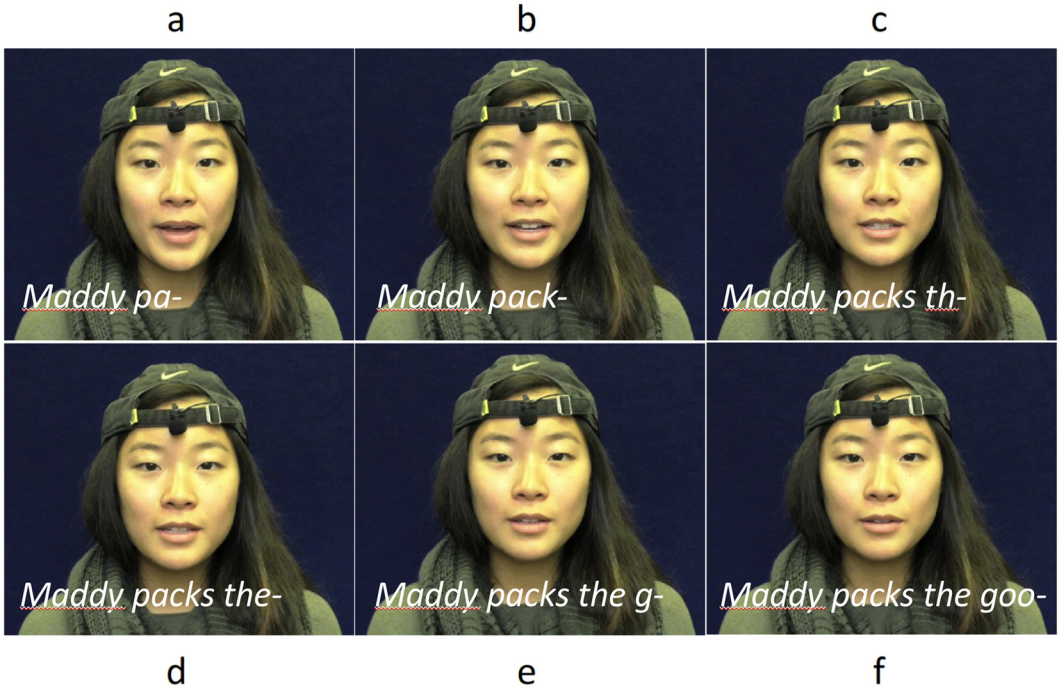

FIG. 2. (Color online) Audio-video recorded sentences were gated at locations that corresponded to specific segmental articulations. Here, the final frame before each of six sentence-internal gates is shown for a single production of "Maddy packs the goop."

C. Perception task

1. Procedure

Perceivers were assigned either to the AV speech condition or to the audio-only speech condition. Within each condition, perceivers were assigned either to the phrasal verb condition or the plain verb condition. Stimuli were blocked by speaker and presented in random order. Each perceiver made judgments on sentences produced by three speakers. The subset of speakers varied across perceivers as did the order in which they were presented to perceivers. The sentences produced by each speaker were judged by a minimum of ten perceivers.

Perceivers were told that the experiment was not about them, but rather that they were serving as our measurement device in a production experiment on coarticulation. The concept of coarticulation was explained. Perceivers were then instructed specifically in the gated speech task and told to do their best to guess whether the sentence-final word rhymed with "oop" or with "ack." They were also explicitly told to treat the "hat" tokens as rhyming with "ack" rather than with "oop." Complete instructions for the AV condition can be found in the Appendix. The instructions were nearly identical in the audio-only condition, except that it was made clear to perceivers that they would only be hearing the short clips rather than watching them.

Once instructed in the task, perceivers sat in front of an iMac computer (Apple) with a 21-inch monitor. The AV stimuli were played in QuickTime (Apple), and the audio-only stimuli in Praat (Boersma and Weenink, 2017). Once played through, two buttons appeared: one with the label "oop" and the other with the label "ack." Perceivers made their guess by clicking on one of the buttons, after which the next stimulus was presented. Since the task included mandatory breaks between the by-speaker blocks, it took perceivers about 45 min to complete judgments on three speakers.

TABLE II. The mean interval in frames between each gate for each sentence type and speaker. Standard deviations are shown in parentheses. Video was recorded at 30 fps, so each frame is equal to an interval of 33 ms.

		Gate interval in frames					
Sentence	Speaker	b-a	c-b	d-c	e-d	f-e	g-f
Phrasal verb	1F	1.70 (0.47)	1.17 (0.38)	1.83 (0.75)	2.70 (0.99)	12.83 (1.63)	
	2F	1.57 (0.50)	1.07 (0.25)	2.03 (0.61)	3.30 (0.65)	8.17 (1.31)	
	3F	1.20 (3.65)	1.00 (0.00)	1.70 (0.60)	3.34 (0.61)	12.97 (1.38)	
	4F	1.97 (0.49)	1.00 (0.18)	1.30 (0.47)	2.33 (2.83)	13.37 (2.83)	
	5F	1.97 (0.41)	1.00 (0.00)	2.10 (0.66)	2.37 (1.73)	12.70 (1.72)	
Plain verb	1F	3.93 (0.52)	4.30 (0.60)	1.70 (0.47)	2.47 (0.78)	2.17 (0.38)	13.77 (2.45)
	2F	3.57 (0.50)	4.07 (0.37)	1.60 (0.50)	2.20 (0.41)	2.70 (0.75)	14.03 (3.08)
	3F	4.13 (1.01)	4.97 (0.93)	1.83 (0.38)	2.37 (0.61)	2.55 (0.69)	17.70 (8.35)
	4F	2.73 (0.52)	3.50 (0.63)	1.40 (0.50)	1.67 (0.48)	2.10 (1.06)	13.40 (2.52)
	5F	3.93 (0.64)	4.13 (0.73)	1.70 (0.47)	2.23 (0.57)	2.33 (0.61)	13.63 (2.08)

2. Outcome measure

The sentences were designed with bilabials to provide highly salient visual landmarks for segmentation purposes. The bilabial gesture occurred one more time and closer to the target determiner phrase in sentences with the phrasal verb (i.e., "*Maddy packs up the*…") than in sentences with the plain verb (i.e., "*Maddy packs the*…"). This design feature may explain why preliminary investigations of the data revealed that perceivers in the AV condition were biased towards an "oop" response when sentences included a phrasal verb, but not when they included a plain verb. To correct for this and any other potential biases, all correct responses were summed across perceivers and repetitions within speaker, condition, verb type, noun onset, and stimulus gate. The summed correct responses were then divided by the sum of all responses at that gate for that noun onset, verb type, condition, and speaker. The resulting proportion correct is the measure of bias-corrected accuracy reported herein.

D. Other measures

1. Vowel acoustics

Acoustic measurement was limited to schwa in the target determiner. Duration, means $F1$ and $F2$, and the slope of $F2$ from midpoint to offset were measured. Segmentation was based on visual inspection of the oscillogram and spectrogram in Praat. Onset and offset boundaries were defined by visible abrupt changes in the oscillogram, periodicity, and the presence of $F2$ formant structure. The vowel formants were tracked using linear predictive coding (LPC) with the maximum number of formants set to five and maximum formant frequency to 5500 Hz. All tracks were visually inspected. If deemed accurate, the corresponding means $F1$ and $F2$ values were automatically extracted over total schwa duration. $F2$ midpoint and offset were also recorded in order to calculate slope. If the LPC tracks were off, they were hand-corrected by adjusting either the maximum number of formants or maximum formant frequency before formant values were extracted.

2. Perioral movement profiles

We used optical flow analysis (Barbosa *et al.*, 2008) to obtain a global measure of speech kinematics. Optical flow analysis was chosen because it could be applied directly to the videos used as stimuli. Moreover, just like the perceiver's judgments, the estimate was synthetic: it combined movement into values that represented summed horizontal and vertical orofacial movement from one frame to the next. Barbosa and colleagues explain the algorithm as follows:

> "Roughly speaking, after conversion to grayscale, the algorithm compares consecutive frames of the video sequence and calculates how much and in which direction each pixel in the image moved from one frame to the next. The algorithm then assigns to each pixel a displacement vector corresponding to the difference in the pixel position across the two frames. The array of displacement vectors comprises the optical flow field" (p. 173).

It is from the optical flow field that frame-to-frame displacement is calculated across all frames in a video clip. More specifically, the optical flow field is the vector velocity of individual pixels within the field, which is calculated using the discrete time distance between frames (1 divided by the frame rate) and the magnitude of displacement of the pixel. The dimensionality of a given optical flow field is equal to the number of rows of pixels times the number of columns. Thus, at a standard resolution of 640×480, the field has 307 200 vectors. In order to reduce the field to a space that is tractable for analysis, the vectors are summed vertically and horizontally to obtain a measure of cumulative two-dimensional (2D) movement from one frame to the next.

Barbosa and colleagues (2008) showed that cumulative movement calculated using optical flow analysis is highly correlated with temporal modulation of the acoustic signal, even though the contribution from individual articulators to speech sounds is lost. They further note that more specific information about segmental articulations can be obtained with the algorithm by defining *regions of interest* and *regions of disinterest*, where information is simply zeroed out. This was the approach we took here. Two regions of interest were defined: a perioral region and a head region. Both were defined in the first frame of each clip associated with each sentence that the speaker produced. The perioral region was defined from the base of the nose to maximal jaw opening in the vertical plane and for the entire jaw region in the horizontal plane. The head region was from the base of the nose to the top of the head in the vertical plane and the full width of the head in the horizontal plane to capture suprasegmental movements of the face and head. Figure 3 illustrates the regions of interest for a single speaker. Note that the regions extend beyond the edge of the speaker's face in order to capture the full extent of vertical and horizontal frame-to-frame movements of the jaw and head during sentence production.

Once cumulative vertical and horizontal flow values were obtained for both regions of interest, values from the

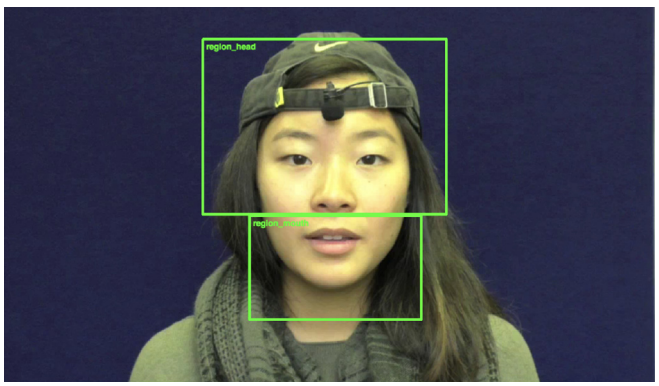

FIG. 3. (Color online) Optical flow analysis was used to measure cumulative frame-to-frame 2D movement in the perioral region. Correlated head movements were regressed out.

TABLE III. Bias-corrected accuracy predicting the target noun rhymes at five different gates in phrasal verb sentences produced by five female speakers. Accuracy values are based on pooled perceiver judgments. Chance performance is 0.50 accuracy.

Speaker	Condition	Onset	"up" C a	"the" C b	"the" V c	Noun C d	Noun V e	Full sentence f
1F	AV	h	0.51	0.59	0.93	0.99	1	1
		g	0.51	0.56	0.84	0.91	0.98	1
		s	0.53	0.59	0.79	0.89	1	1
	Audio-only	h	0.48	0.65	0.89	1	1	1
		g	0.48	0.61	0.76	0.87	1	1
		s	0.47	0.44	0.54	0.63	1	1
2F	AV	h	0.5	0.66	0.93	1	1	1
		g	0.53	0.56	0.78	0.95	1	1
		s	0.53	0.51	0.68	0.89	0.99	1
	Audio-only	h	0.49	0.54	0.86	0.99	0.99	0.99
		g	0.58	0.65	0.75	0.8	0.99	1
		s	0.51	0.44	0.63	0.71	0.99	1
3F	AV	h	0.5	0.83	0.98	0.99	1	1
		g	0.53	0.63	0.83	0.93	0.98	0.98
		s	0.49	0.67	0.78	0.94	1	1
	Audio-only	h	0.51	0.67	0.9	0.99	1	1
		g	0.54	0.56	0.77	0.82	0.98	1
		s	0.61	0.61	0.56	0.8	1	0.99
4F	AV	h	0.5	0.81	0.96	0.98	1	1
		g	0.53	0.68	0.86	0.94	1	1
		s	0.5	0.63	0.64	0.88	1	0.99
	Audio-only	h	0.47	0.62	0.88	0.96	0.99	1
		g	0.48	0.7	0.82	0.87	1	1
		s	0.51	0.54	0.51	0.61	1	1
5F	AV	h	0.51	0.67	0.91	1	1	1
		g	0.44	0.63	0.77	0.96	1	1
		s	0.47	0.58	0.63	0.94	0.99	1
	Audio-only	h	0.45	0.53	0.86	0.98	1	1
		g	0.57	0.58	0.72	0.88	1	1
		s	0.42	0.53	0.5	0.77	1	1
		Mean	0.51	0.61	0.78	0.90	1.00	1.00
		Standard deviation	0.04	0.09	0.14	0.11	0.01	0.00

head region were used to predict values in the perioral region in a linear regression. The results reported below are based on the residualized values from the analysis; that is, on perioral movement with head movement regressed out. This was our synthetic measure of perioral facial kinematics during sentence production.

III. RESULTS AND DISCUSSION

A. Perceiver judgments

Bias-corrected response accuracy is presented in Tables III and IV for each speaker by experimental condition, noun onset, and stimulus gate. Overall response accuracy across speakers is plotted in Fig. 4 by sentence type, experimental condition, and stimulus gate. These response accuracy data indicate that perceivers were unable to predict the target noun rhyme until the onset of the determiner (i.e., gate "b" for phrasal verb sentences and gate "c" for plain verb sentences). Perceivers performed at chance (=0.50 accuracy) when provided with stimuli gated at the final consonant of the verb; specifically, during the /p/ in "packs up" and the /s/ in "packs." Not surprisingly, perceivers performed at ceiling (=1.0 accuracy) when provided with information up until the midpoint of the target vowel in the target noun; specifically, gate "e" for phrasal verb sentences and gate "f" for plain verb sentences. For this reason, we restricted statistical analyses for effects of condition (AV vs audio-only), verb type (phrasal vs plain), and noun onset (/h/ vs /g/ vs /s/) on response accuracy to three gates: the determiner onset, the determiner vowel, and the target noun onset. These gates were treated as a fourth fixed factor in linear mixed effects model analysis, conducted using the lme4 library in R (Bates *et al.*, 2015; R Core Team, 2014). Speaker was the random effect with a random intercept specified. Significant fixed effects were assessed using model comparison. The chi-square statistic associated with difference in model fit is reported here for significant effects. Standard visual diagnostics (normal Q-Q and residuals plots) indicated that neither the assumption of normality nor that of homogeneity of variances were violated.

The analysis indicated a significant four-way interaction ($\chi^2 = 121.02$, $p < 0.01$). All three-way interactions were also significant. Figures 5 and 6 show the two strongest of these: the interactions between experimental condition, noun onset, and gate ($\chi^2 = 85.71$, $p < 0.01$) and between sentence type, noun onset, and gate ($\chi^2 = 42.64$, $p < 0.01$). The weakest

TABLE IV. Bias-corrected accuracy in predicting the target noun rhymes at six different gates in plain verb sentences produced by five female speakers. Accuracy values are based on pooled perceiver judgments. Chance performance is 0.50 accuracy.

Speaker	Condition	Onset	"pack" V a	"pack" C b	"the" C c	"the" V d	Noun C e	Noun V f	Full sentence g
1F	AV	h	0.51	0.53	0.8	0.98	0.99	1	1
		g	0.55	0.54	0.73	0.94	1	1	1
		s	0.51	0.57	0.78	0.89	0.98	0.99	1
	Audio-only	h	0.47	0.5	0.55	0.92	0.99	1	0.99
		g	0.49	0.52	0.56	0.7	0.94	1	1
		s	0.5	0.54	0.47	0.52	0.75	0.99	1
2F	AV	h	0.53	0.51	0.7	0.92	1	1	0.98
		g	0.51	0.43	0.64	0.81	0.98	0.99	1
		s	0.53	0.48	0.72	0.84	0.94	1	0.99
	Audio-only	h	0.54	0.49	0.52	0.84	0.98	0.99	1
		g	0.49	0.51	0.56	0.77	0.93	0.99	0.99
		s	0.46	0.48	0.51	0.63	0.71	0.99	1
3F	AV	h	0.48	0.55	0.73	0.97	1	1	1
		g	0.48	0.49	0.72	0.89	1	1	1
		s	0.43	0.5	0.7	0.85	0.97	1	1
	Audio-only	h	0.53	0.53	0.53	0.83	0.98	1	0.98
		g	0.5	0.49	0.5	0.66	0.94	1	1
		s	0.56	0.48	0.54	0.58	0.7	1	1
4F	AV	h	0.58	0.58	0.82	0.96	0.99	0.98	1
		g	0.5	0.53	0.76	0.89	0.94	0.97	1
		s	0.47	0.48	0.67	0.74	0.86	1	0.97
	Audio-only	h	0.51	0.53	0.44	0.9	0.99	0.96	0.97
		g	0.54	0.56	0.51	0.74	0.89	0.96	1
		s	0.5	0.43	0.4	0.53	0.49	1	0.99
5F	AV	h	0.48	0.52	0.69	0.9	1	0.99	1
		g	0.53	0.5	0.54	0.84	1	0.98	1
		s	0.39	0.5	0.48	0.6	0.86	0.99	1
	Audio-only	h	0.53	0.49	0.55	0.8	0.98	0.99	0.98
		g	0.49	0.55	0.58	0.83	0.93	0.99	0.99
		s	0.51	0.48	0.55	0.51	0.75	0.99	0.99
		Mean	0.50	0.51	0.61	0.79	0.92	0.99	0.99
		Standard deviation	0.04	0.04	0.12	0.14	0.12	0.01	0.01

three-way interaction did not include noun onset and was instead between experimental condition, sentence type, and gate ($\chi^2 = 14.21$, $p = 0.04$). Other significant interactions were between experimental condition and noun onset ($\chi^2 = 20.21$, $p < 0.01$), noun onset and gate ($\chi^2 = 35.43$, $p < 0.01$), and between experimental condition and sentence type ($\chi^2 = 8.01$, $p < 0.01$). These interactions are also evident in Figs. 5 and 6. Finally, the simple effects of experimental condition, noun onset, and gate were significant (condition,

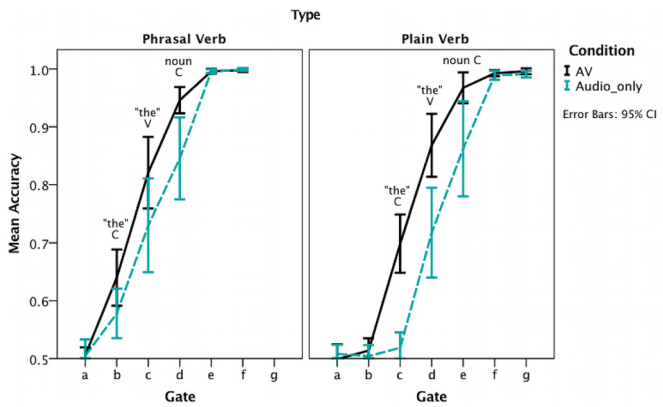

FIG. 4. (Color online) Bias-corrected accuracy in predicting the target noun rhymes in the different sentence types is averaged across speakers and shown by experimental condition and gate. Gates "f" and "g" are equivalent to the whole sentence in the phrasal verb and plain verb sentences, respectively. Accuracy of 0.50 equals chance performance. Error bars indicate the 95% confidence interval.

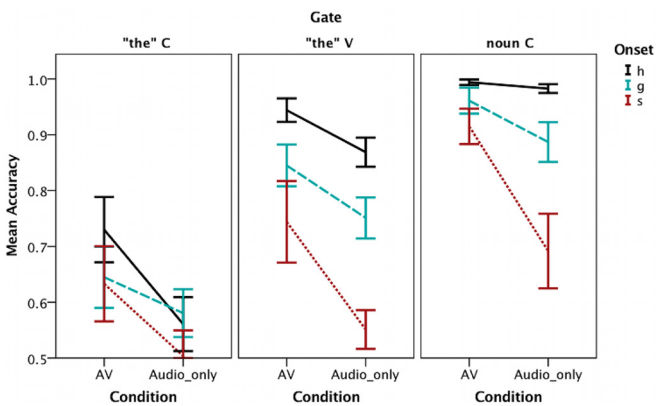

FIG. 5. (Color online) Effects of experimental condition, target noun onset (solid line = \h\; dashed line = \g\; dotted line = \s\) and gate (limited to the three shown) on bias-corrected prediction accuracy averaged across sentence types and speakers. Accuracy of 0.50 equals chance performance. Error bars indicate the 95% confidence interval.

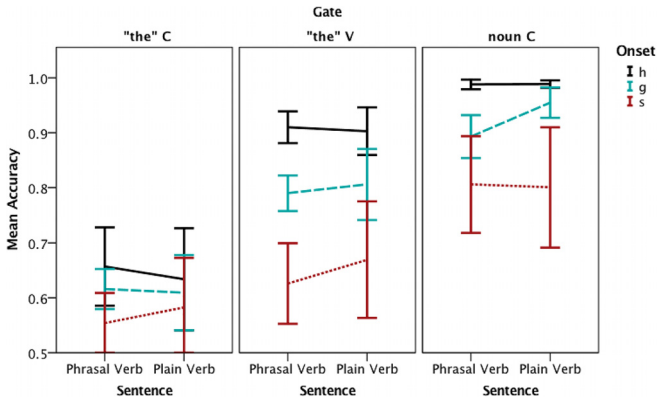

FIG. 6. (Color online) Effects of sentence type (=verb), target noun onset (solid line = \h\; dashed line = \g\; dotted line = \s\) and gate (limited to the three shown) on bias-corrected prediction accuracy is shown averaged across experimental conditions and speakers. Accuracy of 0.50 equals chance performance. Error bars indicate the 95% confidence interval.

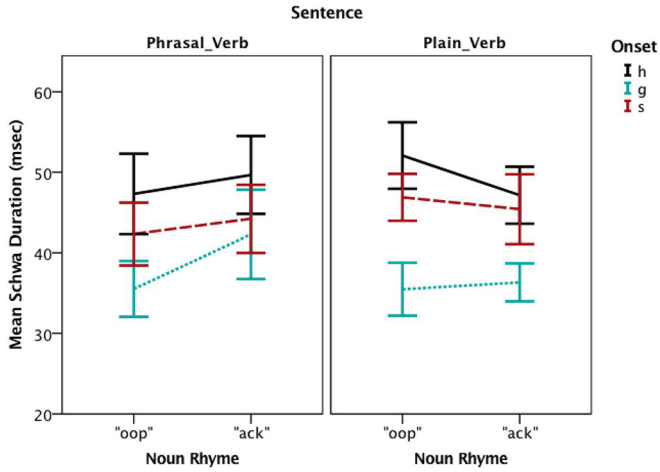

FIG. 7. (Color online) Effect of sentence type (verb), noun onset (solid line = \h\; dashed line = \g\; dotted line = \s\) and rhyme on schwa $F1$ (bottom) and $F2$ (top) in the determiner that modified the "oop" and "ack" nouns. Error bars indicate the 95% confidence interval.

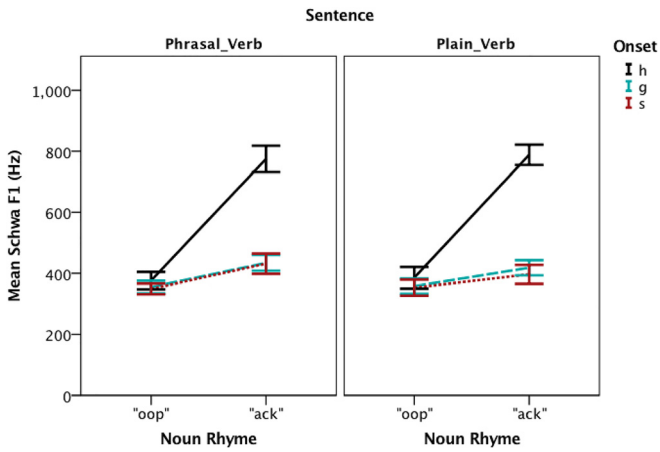

FIG. 8. (Color online) Effect of sentence type (verb), noun onset (solid line = \h\; dashed line = \g\; dotted line = \s\) and rhyme on schwa $F1$ in the determiner that modified the "oop" and "ack" nouns. Error bars indicate the 95% confidence interval.

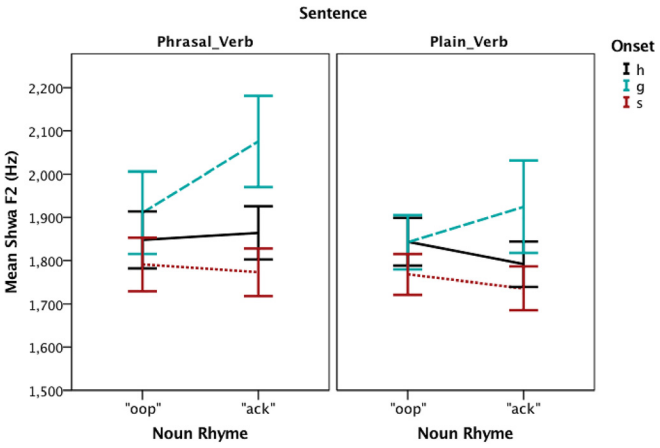

FIG. 9. (Color online) Effect of sentence type (verb), noun onset (solid line = \h\; dashed line = \g\; dotted line = \s\) and rhyme on schwa $F2$ in the determiner that modified the "oop" and "ack" nouns. Error bars indicate the 95% confidence interval.

$\chi^2 = 86.47$, $p < 0.01$; onset, $\chi^2 = 120.38$, $p < 0.01$; gate, $\chi^2 = 238.33$, $p < 0.01$). The effect of sentence type was not significant nor did it interact by itself with noun onset or gate: anticipatory posturing for the "oop" rhyme occurred at the onset of the determiner no matter the verb. This result suggests that word boundaries are relevant for defining the domain of anticipatory coarticulation even if metrical boundaries are not.

The effect of the three gates of interest on response accuracy was straightforward and in line with predictions. Figures 5 and 6 both show that perceivers were better able to predict the identity of that rhyme as the gates approached the target noun rhyme. The effect of condition was also as predicted. Figures 4 and 5 both clearly show that perceivers in the AV speech condition performed better than those in the audio-only condition. Specifically, perceivers in the AV condition performed at well above chance levels in stimuli gated at the onset of the determiner; perceivers in the audio-only condition did not. Both groups of perceivers were at above chance levels when stimuli were gated during the production of the determiner vowel, but even at this gate perceiver performance in the AV speech condition was better overall than performance in the audio-only condition. These results are consistent with the robust finding that AV speech processing is superior to audio-only speech processing (e.g., MacLeod and Summerfield, 1987; Ross *et al.*, 2006; Summerfield, 1992), which underscores the importance of using AV speech instead of audio-only speech for obtaining perceptual measures of coarticulation. Finally, the effect of target noun onset was consistent with differences in coarticulatory resistance. Figures 5 and 6 both show that overall accuracy in predicting rhymes was lowest when nouns began with /s/ ($\beta = -0.10$), particularly in the audio-only condition, and highest when the nouns began with /h/ ($\beta = 0.07$). Even the effect of experimental condition disappeared in stimuli gated at noun onset when this was /h/ (see rightmost panel in Fig. 5).

B. Schwa acoustics

Linear mixed effects models were used to investigate effects of sentence type, noun onset, and noun rhyme on the

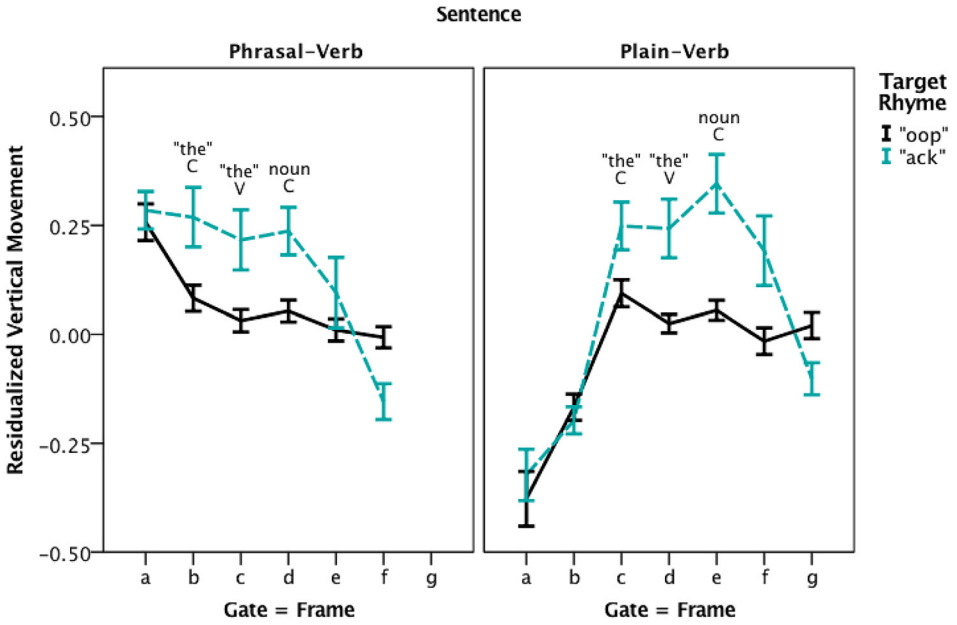

FIG. 10. (Color online) Summed vertical movement in the perioral region with head movement removed. The data are shown as a function of sentence type and the final frame of each of the forward gated stimuli sentences. Frames associated with gates "f" and "g" are equivalent to the whole sentence in the phrasal verb and plain verb sentences, respectively. Error bars indicate the 95% confidence interval.

mean duration, $F1$, $F2$, and $F2$ slope to further validate the perceptual measures of anticipatory coarticulation. In so far as perceivers in the audio-only condition only had access to speech acoustics, these acoustic analyses also help to explain the strong effect of noun onset on mean accuracy in the audio-only condition.

The analyses indicated a significant three-way interaction on schwa duration ($\chi^2 = 18.69$, $p < 0.01$), $F1$ ($\chi^2 = 312.07$, $p < 0.01$), and $F2$ ($\chi^2 = 38.83$, $p < 0.01$), as shown in Figs. 7, 8, and 9, respectively. The two-way interaction between sentence type and noun rhyme was also significant for schwa duration ($\chi^2 = 7.29$, $p < 0.01$) and $F2$ ($\chi^2 = 4.21$, $p = 0.04$). The interaction between sentence type and noun onset was significant for $F2$ only ($\chi^2 = 6.20$, $p = 0.04$). In contrast, the interaction between noun onset and noun rhyme was significant for both $F1$ ($\chi^2 = 306.71$, $p < 0.01$) and $F2$ ($\chi^2 = 26.56$, $p < 0.01$). Finally, the simple effect of sentence type was significant for $F2$ only ($\chi^2 = 21.82$, $p < 0.01$) and that of noun rhyme was significant for $F1$ only ($\chi^2 = 175.89$, $p < 0.01$), but the simple effect of noun onset was significant on all measures taken (schwa duration: $\chi^2 = 78.81$, $p < 0.01$; $F1$: $\chi^2 = 183.64$, $p < 0.01$; $F2$: $\chi^2 = 81.79$, $p < 0.01$; $F2$ slope: $\chi^2 = 20.89$, $p < 0.01$).

Taken together, the results indicate that all measures of schwa acoustics varied especially with noun onset. Figure 7 shows that schwa was longest before /h/ and shortest before /g/. Noun rhyme also influenced schwa acoustics, but the extent to which it did varied substantially with noun onset. Figure 8 shows that $F1$ was dramatically higher before "ack" than before "oop" when the noun onset was /h/, but only somewhat higher when the noun onset was either /g/ or /s/. Figure 9 shows that $F2$ was also higher before "ack" than before "oop," but only when noun onset was /g/. The strong interaction with noun onset in this case could suggest an allophonic difference in /g/ production as a function of the rhyme. Overall, the results help to explain why prediction accuracy was especially influenced by target noun onsets in the audio-only condition where perceivers only had access to acoustic information.

C. Perioral movement profiles

Perceivers in the AV condition had access to visual information in addition to acoustic information. Movement profiles derived from the optical flow analysis show that they made good use of visual information in the gated speech task. The horizontal and vertical movement profiles shown in Fig. 10 suggests that most of the information about coarticulation was derived from the speaker's movement in the vertical plane. In particular, the vertical movement profiles indicate that speakers produced sentences with final "oop" and "ack" rhymes differently from the outset of the determiner no matter the verb. This pattern provides an explanation for why perceivers in the AV speech condition were able to predict the noun rhyme at above chance levels before hearing the acoustic correlates of anticipatory posturing in the determiner vowel (see Figs. 4, 5, and 6).

Figure 10 further suggests that frame-to-frame movement differences increased over the next several gates before beginning to converge at vowel midpoint in the noun (gate "e" for phrasal verbs, gate "f" for plain verbs). This pattern suggests that visual information helps to reinforce acoustic information available in the vowel and so explains why perceivers in the AV condition outperformed those in the audio-only condition. Finally, frame-to-frame movement displacement values were larger in anticipation of the "ack" rhyme than in anticipation of the "oop" rhyme across most gates, which is as expected given the different vowel targets.

To further investigate parallels between the measurement of vertical movement in the AV clips and response accuracy in the gated AV speech task, we computed a measure that could be analyzed by verb type, noun onset, and gate. This measure was the absolute difference between sentences with final "oop" and "ack" in residualized horizontal and vertical displacement values. Differences were

calculated within speaker at the frames associated with the determiner onset, determiner vowel, and noun onset gates. Recall that it is only at these gates that we found effects of verb type and noun onset on response accuracy in the AV speech condition; response accuracy was at chance before the determiner onset and at ceiling after the noun onset.

Mixed effects modeling confirmed the absence of a difference between the horizontal movement profiles; none of the effects were significant. In contrast, the absolute difference values in the vertical movement profiles varied systematically with the three-way interaction between sentence type, noun onset, and gate ($\chi^2 = 74.85$, $p < 0.01$). The two-way interaction between noun onset and gate was also significant ($\chi^2 = 63.10$, $p < 0.01$), as was the simple effect of noun onset ($\chi^2 = 44.65$, $p < 0.01$). The three-way results are shown in Fig. 11.

The largest differences between the minimal pair sentences varied by gate and target noun. When the noun began with /h/ (i.e., "hoop" vs "hat") differences were largest at the onset of the determiner ("the" C) and determiner vowel ("the" V). When nouns began with /g/ (i.e., "goop" vs "gak") differences were largest at the gate that corresponded to production of the noun onset itself. Note that at this gate, the absolute difference in cumulative vertical movement for "hoop" vs "hat" decreased. This is very likely because speakers attained near target position for the upcoming rhyme during the determiner itself and so did not need to reposition the articulators at noun onset. Note also that absolute movement differences between the sentences with final "oop" and "ack" were smallest when the noun began with /s/ (i.e., "soup" vs "sack"). This result is consistent with a high degree of coarticulatory resistance, resulting in greater anticipatory posturing for /s/ rather than for the "oop" or "ack" rhymes. Altogether, the pattern of results shown in Fig. 11 provide a basis for the effect of onset on response accuracy in the AV speech condition: perceivers were best able to predict an "oop" or "ack" rhyme when the noun began with /h/ and least able to do so when it began with /s/.

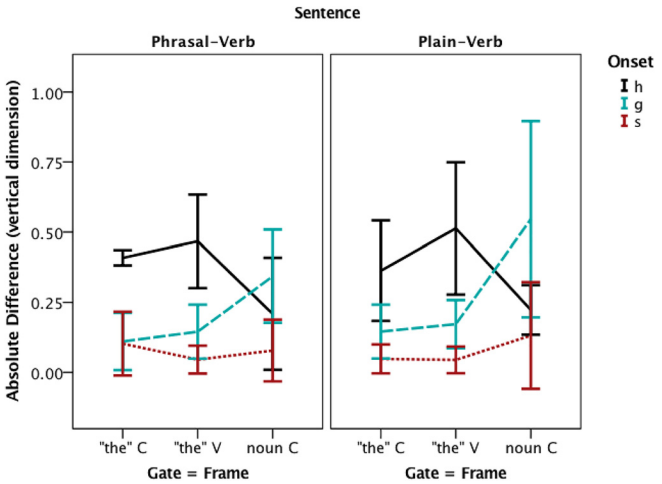

FIG. 11. (Color online) Effects of sentence type, noun onset (solid line =\h\; dashed line =\g\; dotted line =\s\) and gate (limited to the three shown) on the absolute difference in perioral movement for sentences ending in "oop" and "ack" rhymes. Error bars indicate the 95% confidence interval.

IV. GENERAL DISCUSSION

The goal of the study was to introduce a method for measuring long-distance coarticulation that can be used to investigate the psychology of speech production across different populations, including children and other individuals with low inhibitory control. The method leverages work in AV speech perception to assess degree of coarticulation at experimentally defined temporal locations using gated speech. The study aim was to validate the method. This aim was met. Perceivers' ability to predict the identity of an upcoming rhyme conformed both to expected effects of distance from target and expected articulatory constraints on coarticulation. Moreover, the perceptual results were clearly in line with acoustic and kinematic measurement of the stimuli.

The study demonstrated how the method can be used for testing specific hypotheses arising from psychological models of speech production. For example, the current stimuli were designed to test the hypothesis that metrical structure defines production units (e.g., Roelofs and Meyer, 1998) and therefore the domain of articulatory preplanning (e.g., Jordan, 1997). The metrical context for the determiner was manipulated by varying the phonological structure of the verb. Following metrical theory (Hayes, 1982), this manipulation should have resulted in a metrical boundary after "the" in the plain verb context. In the phrasal verb context, the metrical boundary should be between the verb and the determiner, leaving "the" unfooted. The sentences were gated to test for a boundary effect. The predictions were that response accuracy would be above chance at determiner onset in the phrasal verb context, but not in the plain verb context. Contrary to this prediction, the effect of sentence type did not by itself have a significant effect on accuracy nor did it interact with gate. This finding suggests that a metrical boundary has very little effect on the domain of coarticulation. Our conclusion based on the study results extends Magen's (1997) original conclusion to the supra-lexical context.

Although the metrical boundary hypothesis was not supported, Fig. 4 suggests a difference in coarticulatory strength across sentence types that is informative from a methodological perspective. In particular, the somewhat poorer response accuracy at the determiner onset gate in the phrasal verb sentences (gate "b") compared to the plain verb sentences (gate "c") was driven by a stronger bias toward an "oop" response in the former condition compared to the latter. This bias was likely due to the fact that the first gate for phrasal verb sentences was introduced at the moment of bilabial closure during "up" production, which may have rendered labial articulation especially salient at the onset of "the" production. Examination of the inter-gate intervals shown in Table II is consistent with this interpretation in that the onset of "the" also occurred closer in time to the offset of the verb in the phrasal verb condition [b-a interval, $M = 1.68$ frames, standard deviation (SD) $= 0.32$ frames] compared to the plain verb condition (c-b interval, $M = 4.19$ frames, $SD = 0.53$ frames). Importantly, the interval between the

onset of "the" and vowel midpoint of the target rhyme did not differ between sentence conditions (e-b interval, $M = 5.65$ frames, $SD = 1.26$ frames vs f-c interval, $M = 5.95$ frames, $SD = 3.16$ frames). Overall, the apparent unintended effect of using "up" as the phrasal verb particle on both perceiver performance and inter-gate intervals suggests that some caution must be taken when creating the gated AV stimuli that are central to our measurement method.

Another goal of the present study was to show that AV speech perception is superior to auditory-only perception for measuring the scope and strength of anticipatory coarticulation. This goal was also met. Response accuracy was higher overall in the AV speech condition than in the audio-only speech condition. Perceivers in the AV speech condition were also able to predict an upcoming rhyme earlier than in the audio-only speech condition. These expected results can be explained with reference to the complementary acoustic and kinematic analyses. The kinematic analyses indicated that information about speaker adjustments for an upcoming rhyme was available earlier in the AV speech condition than in the audio-only condition. For example, movement differences in anticipation of "oop" vs "ack" were evident in the video from the onset of "the."

In light of the parallelisms between the perceptually derived results and the kinematic results, one might question why we have bothered developing a perceptually based method for measuring anticipatory coarticulation when the equally noninvasive method of optical flow analysis is available. This question is important given that optical flow analysis is much less resource consuming than the gated AV speech method we have developed: the gated AV speech method requires both time-consuming video-editing and running a large number of human subjects as perceivers; optical flow analysis requires only that the user identify intervals of interest in the video-recorded speech stream.

The principle reason to use a perceptually based measurement of anticipatory coarticulation is that it is more robust for use with wiggly children than optical flow analysis. Optical flow analysis is limited to the information provided in 2D video frames. All pixels in the 2D array are treated equally, and all frame-to-frame changes in individual pixel light characteristics are noted and summed. This not only means that larger movements contribute more to the sum than smaller movements, but that any change in speaker orientation with respect to the camera changes the relative contribution of different speech-related movements to the sum. In the present study, we focused on perioral movement using optical flow analysis by also measuring and then regressing out head movement. This fix was sufficient to provide the clearly patterned results discussed above, but it was only possible because our speakers looked directly into the camcorder while speaking and moved very little. Had they shifted their head away from midline or looked down while speaking, and so diminished or otherwise altered the perspective of the perioral region in the 2D plane, head movements would have swamped out other speech-related movements in the resulting summed movement profiles.

In contrast to measurements that are limited by the format of the recorded data, measurements provided by human perceivers are enhanced by the knowledge humans bring to the measurement task. Importantly for our purposes, perceiver knowledge results in automatic normalization across speakers and changes in the visual scene. For example, Jordan and colleagues have demonstrated that perceiver performance in AV speech perception tasks is insensitive to the size of the speaker's image on the screen (Jordan and Sergeant, 1998), to the speaker's orientation with respect to the camera (Jordan *et al.*, 1997; Jordan and Thomas, 2001),[2] and, surprisingly, even to whether or not the speaker's face is upside-down in the frame (Jordan and Bevan, 1997). This ability to automatically adjust for size and perspective implies that perceiver judgments of speech-related movements are highly robust in the face of natural movement during speaking. It is this robustness that especially recommends the method for studying anticipatory coarticulation in children, who have a much more difficult time than adults staying stock-still while completing speech tasks in a laboratory setting.

V. CONCLUSION

To summarize, we conducted a two-part study to validate a perceptually based method that uses prediction accuracy in gated AV speech to measure the scope and strength of anticipatory coarticulation at experimentally defined locations in the speech stream. Results showed that AV speech perception is superior to auditory-only speech perception for detecting coarticulatory effects. Prediction accuracy was above chance at an earlier gate in the AV condition compared to the audio-only condition. The gate in question was the onset of "the" production, which was selected to test for an effect of metrical structure on coarticulation—highlighting a particular strength of the method: it affords a method for testing specific predictions regarding the onset of anticipatory coarticulation. Prediction accuracy was also higher overall in the AV condition compared to the audio-only condition, suggesting the importance of speech kinematics for detecting coarticulation. Finally, prediction accuracy varied systematically with the target noun onset, but not with metrical context. These results are consistent with well-known effects of coarticulatory resistance on the scope and strength of anticipatory effects, and with Magen's (1997) conclusion that coarticulation is insensitive to metrical boundaries. Results from acoustic and kinematic analyses of the stimuli supported the conclusion that prediction accuracy in the gated AV speech task tracks speech production patterns.

Going forward we will apply our method to the study of long-distance anticipatory coarticulation in children and other populations that may tolerate less well than typical adults the sensors and wires, probes, and movement constraints imposed by commonly used methods for kinematic measurement. The results clearly show that the gated AV speech method can be used when research questions are about underlying plan structure or the psychology of speech production more generally. More invasive methods for measuring speech kinematics are still required to study the detailed mechanics of articulation.

ACKNOWLEDGMENTS

E.V.-B. (1952–2017) contributed substantially to this study, but passed away before the manuscript was written. Any problems with the analyses or interpretation of results presented herein should be attributed to the other authors. The work reported herein was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) under Grant No. R01HD087452 (PI: Redford). The content is solely the authors' responsibility and does not necessarily reflect the views of NICHD.

APPENDIX: INSTRUCTIONS FOR PERCEPTION TASK

We are not studying you, we are studying the speakers you will be seeing. You are our measurement devices. We are measuring anticipatory coarticulation, which is the implicit planning you do in your head to move your speech articulators in advance of the sound you will make. 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." The speakers you will be listening to are saying sentences that end in an "oop" or "ack" word. The sentences are (experimenter lists sentences here and notes that sentences ending with "hat" are to be categorized with the "ack"-ending sentences). You will see clips of them saying the sentences, but the clips are sometimes cut fairly short so that you have less information from which to make the judgment. Just do your best and try to decide whether the speaker is going to be saying the sentence with a final "oop" or "ack" word.

[1]Legacy versions of Final Cut Pro cannot be run on any of the most recent operating systems. Final Cut Pro X provides only a rectified audio waveform that also has less temporal detail than in the legacy versions. Accordingly, we used Lightworks Pro (Editshare, Watertown, MA) to create Fig. 1.

[2]A reviewer notes that a three-quarter speaker orientation may be better than other orientations for viewers to perceive rounding in a language like French, where lip protrusion is more prominent during rounding than in English (e.g., Noiray *et al.*, 2008).

Barbosa, A. L., Yehia, H. C., and Vatikiotis-Bateson, E. (**2008**). "Linguistically valid movement behavior measured non-invasively," in *Proceedings of the International Conference on Auditory-Visual Speech Processing—AVSP 2008*, September, Tangalooma, Australia, pp. 173–177.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Sofware **67**(1), 1–48.

Bell-Berti, F., and Harris, K. S. (**1979**). "Anticipatory coarticulation: Some implications from a study of lip rounding," J. Acoust. Soc. Am. **65**(5), 1268–1270.

Bladon, R. A. W., and Al-Bamerni, A. (**1976**). "Coarticulation resistance in English /l/," J. Phon. **4**(2), 137–150.

Bladon, R. A. W., and Nolan, F. (**1977**). "A video-fluorographic investigation of tip and blade alveolars in English," J. Phon. **5**(2), 185–193.

Boersma, P., and Weenink, D. (**2017**). "Praat: Doing phonetics by computer [computer program]," available at http://www.praat.org/.

Cathiard, M. A., Lallouache, M. T., and Abry, C. (**1996**). "Does movement on the lips mean movement in the mind?," in *Speechreading by Humans and Machines*, edited by D. Stork and M. E. Hennecke (Springer, Berlin), pp. 211–219.

Daniloff, R., and Moll, K. (**1968**). "Coarticulation of lip rounding," J. Speech Hear. Res. **11**(4), 707–721.

Fuchs, S., Petrone, C., Krivokapić, J., and Hoole, P. (**2013**). "Acoustic and respiratory evidence for utterance planning in German," J. Phon. **41**(1), 29–47.

Goffman, L., and Smith, A. (**1999**). "Development and phonetic differentiation of speech movement patterns," J. Exp. Psychol., Hum. Percept. Perform. **25**(3), 649–660.

Green, J. R., Moore, C. A., Higashikawa, M., and Steeve, R. W. (**2000**). "The physiologic development of speech motor control: Lip and jaw coordination," J. Speech, Lang. Hear. Res. **43**(1), 239–255.

Grosvald, M. (**2009**). "Interspeaker variation in the extent and perception of long-distance vowel-to-vowel coarticulation," J. Phon. **37**(2), 173–188.

Hadar, U., Steiner, T. J., Grant, E. C., and Rose, F. C. (**1983**). "Kinematics of head movements accompanying speech during conversation," Hum. Mov. Sci. **2**(1), 35–46.

Hayes, B. (**1982**). "Extrametricality and English stress," Linguist. Inquiry **13**(2), 227–276.

Huber, J. E. (**2008**). "Effects of utterance length and vocal loudness on speech breathing in older adults," Respir. Physiol. Neurobiol. **164**(3), 323–330.

Jordan, M. I. (**1997**). "Serial order: A parallel distributed processing approach," Adv. Psychol. **121**, 471–495.

Jordan, T. R., and Bevan, K. M. (**1997**). "Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition," J. Exp. Psychol., Hum. Percept. Perform. **23**, 388–403.

Jordan, T. R., and Sergeant, P. C. (**1998**). "Effects of facial image size on visual and audiovisual speech recognition," in *Hearing by Eye II: Advances in the Psychology of Speech Reading and Audio-Visual Speech*, edited by R. Campbell, B. Dodd, and D. K. Burnham (Psychology Press, Hove, UK), pp. 155–176.

Jordan, T. R., Sergeant, P. C., Martin, C., Thomas, S. M., and Thow, E. (**1997**). "Effects of horizontal viewing angle on visual and audiovisual speech perception," in *IEEE International Conference on Systems, Man, and Cybernetics* (Institute of Electrical and Electronics Engineers, Piscataway, NJ), Vol. 2, pp. 1626–1631.

Jordan, T. R., and Thomas, S. M. (**2001**). "Effects of horizontal viewing angle on visual and audiovisual speech recognition," J. Exp. Psychol.: Hum. Percept. Perform. **27**(6), 1386–1403.

Katz, W. F., Kripke, C., and Tallal, P. (**1991**). "Anticipatory coarticulation in the speech of adults and young children: Acoustic, perceptual, and video data," J. Speech, Lang. Hear. Res. **34**(6), 1222–1232.

Keating, P. A., Lindblom, B., Lubker, J., and Kreiman, J. (**1994**). "Variability in jaw height for segments in English and Swedish VCVs," J. Phon. **22**(4), 407–422.

Kelso, J. S., Tuller, B., Vatikiotis-Bateson, E., and Fowler, C. A. (**1984**). "Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures," J. Exp. Psychol., Hum. Percept. Perform. **10**(6), 812–832.

Kozhevnikov, A., and Chistovich, L. (**1965**). *Rech: Artikulyatsia i Vospriyatiye* (Nauka, Moscow) *[Speech: Articulation and Perception* (Joint Publications Research Service, U.S. Department of Commerce Translation, Washington, DC)], No. 30, 543.

MacLeod, A., and Summerfield, Q. (**1987**). "Quantifying the contribution of vision to speech perception in noise," Br. J. Audiol. **21**(2), 131–141.

Magen, H. S. (**1997**). "The extent of vowel-to-vowel coarticulation in English," J. Phon. **25**(2), 187–205.

McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.

Moradi, S., Lidestam, B., and Rönnberg, J. (**2013**). "Gated audiovisual speech identification in silence vs. noise: Effects on time and accuracy," Front. Psychol. **4**, 359.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Vatikiotis-Bateson, E. (**2004**). "Visual prosody and speech intelligibility: Head movement improves auditory speech perception," Psychol. Sci. **15**(2), 133–137.

Munhall, K. G., and Tohkura, Y. (**1998**). "Audiovisual gating and the time course of speech perception," J. Acoust. Soc. Am. **104**(1), 530–539.

Munhall, K. G., and Vatikiotis-Bateson, E. (**1998**). "The moving face during speech communication," in *Hearing by Eye II: Advances in the Psychology of Speech Reading and Audio-Visual Speech*, edited by R. Campbell, B. Dodd, and D. K. Burnham (Psychology Press, Hove, UK), pp. 123–139.

Noiray, A., Cathiard, M. A., Abry, C., Ménard, L., and Savariaux, C. (**2008**). "Emergence of a vowel gesture control: Attunement of the anticipatory rounding temporal pattern in French children," in *Emergence of*

Linguistic Abilities, edited by S. Kern, F. Gayraud, and E. Marsico (Cambridge Scholars, Newcastle upon Tyne, UK), pp. 100–117.
Noiray, A., Ménard, L., and Iskarous, K. (**2013**). "The development of motor synergies in children: Ultrasound and acoustic measurements," J. Acoust. Soc. Am. **133**(1), 444–452.
Öhman, S. E. (**1966**). "Coarticulation in VCV utterances: Spectrographic measurements," J. Acoust. Soc. Am. **39**(1), 151–168.
Perkell, J. S., and Klatt, D. H. (eds.). (**1986**). *Invariance and Variability in Speech Processes* (Psychology Press, London, UK).
R Core Team (**2014**). *R: A language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna).
Recasens, D. (**1984**). "Vowel-to-vowel coarticulation in Catalan VCV sequences," J. Acoust. Soc. Am. **76**(6), 1624–1635.
Recasens, D., Pallarès, M. D., and Fontdevila, J. (**1997**). "A model of lingual coarticulation based on articulatory constraints," J. Acoust. Soc. Am. **102**(1), 544–561.
Roelofs, A., and Meyer, A. S. (**1998**). "Metrical structure in planning the production of spoken words," J. Exp. Psychol. Learn. Mem. Cog. **24**(4), 922–939.
Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (**2006**). "Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments," Cereb. Cortex **17**(5), 1147–1153.
Savariaux, C., Perrier, P., and Orliaguet, J. P. (**1995**). "Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production," J. Acoust. Soc. Am. **98**(5), 2428–2442.
Smeele, P. M. T. (**1994**). "Perceiving speech: Integrating auditory and visual speech," Doctoral dissertation, Delft University of Technology.
Stone, M. (**2010**). "Laboratory techniques for investigating speech articulation," in *The Handbook of Phonetic Sciences*, 2nd ed., edited by W. J. Hardcastle, J. Laver, and F. E. Gibbon (Blackwell Publishing Ltd., Hoboken, NJ).
Summerfield, Q. (**1992**). "Lipreading and audio-visual speech perception," Philos. Trans. R. Soc. Lond. B **335**, 71–78.
Thomas, S. M., and Jordan, T. R. (**2004**). "Contributions of oral and extraoral facial movement to visual and audiovisual speech perception," J. Exp. Psychol., Hum. Percept. Perform. **30**(5), 873–888.
Tremblay, S., Shiller, D. M., and Ostry, D. J. (**2003**). "Somatosensory basis of speech production," Nature **423**(6942), 866–869.
Whalen, D. H. (**1990**). "Coarticulation is largely planned," J. Phon. **18**, 3–35.
Winkworth, A. L., Davis, P. J., Adams, R. D., and Ellis, E. (**1995**). "Breathing patterns during spontaneous speech," J. Speech, Lang. Hear. Res. **38**(1), 124–144.
Yehia, H. C., Kuratate, T., and Vatikiotis-Bateson, E. (**2002**). "Linking facial animation, head motion and speech acoustics," J. Phon. **30**(3), 555–568.
Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (**1998**). "Quantitative association of vocal-tract and facial behavior," Speech Commun. **26**(1), 23–43.
Zharkova, N., Hewlett, N., and Hardcastle, W. J. (**2011**). "Coarticulation as an indicator of speech motor control development in children: An ultrasound study," Motor Control **15**(1), 118–140.