Journal of Phonetics 56 (2016) 15–37

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Research Article

Differences in coda voicing trigger changes in gestural timing: A test case from the American English diphthong /aɪ/

Anne Pycha [a,*], Delphine Dahan [b]

[a] Department of Linguistics, University of Wisconsin, Milwaukee, P.O. Box 413, Milwaukee, WI 53201-0413, USA
[b] Department of Psychology, University of Pennsylvania, 3401 Walnut Street, Room 412, Philadelphia, PA 19104-6228, USA

ARTICLE INFO

Article history:
Received 31 March 2014
Received in revised form
5 January 2016
Accepted 12 January 2016

Keywords:
Diphthong
Duration
Gesture
Coda voicing
Articulatory phonology

ABSTRACT

We investigate the hypothesis that duration and spectral differences in vowels before voiceless versus voiced codas originate from a single source, namely the reorganization of articulatory gestures relative to one another in time. As a test case, we examine the American English diphthong /aɪ/, in which the acoustic manifestations of the nucleus /a/ and offglide /ɪ/ gestures are relatively easy to identify, and we use the ratio of nucleus-to-offglide duration as an index of the temporal distance between these gestures. Experiment 1 demonstrates that, in production, the ratio is smaller before voiceless codas than before voiced codas; this effect is consistent across speakers as well as changes in speech rate and phrasal position. Experiment 2 demonstrates that, in perception, diphthongs with contextually incongruent ratios delay listeners' identification of target words containing voiceless codas, even when the other durational and spectral correlates of voicing remain intact. This, we argue, is evidence that listeners are sensitive to the gestural origins of voicing differences. Both sets of results support the idea that the voicing contrast triggers changes in timing: gestures are close to one another in time before voiceless codas, but separated from one another before voiced codas.

1. Introduction

American English vowels undergo several dynamic changes conditioned by voiceless versus voiced coda environments. One obvious change concerns duration. Vowels exhibit consistently shorter durations before voiceless codas than before voiced ones (e.g. House & Fairbanks, 1953; Peterson & Lehiste, 1960). Another change concerns spectral quality. Vowels exhibit more peripheral formant values before voiceless codas, compared to more centralized values before voiced codas (Moreton, 2004 and references cited therein). The puzzle is that these two changes appear disconnected from one another, because when vowels shorten in most other conditioning environments, we typically see centralization, not peripheralization. In this paper, we use concepts from articulatory phonology (Browman & Goldstein, 1992; Gafos, 2002; Goldstein & Fowler, 2003) to investigate how the duration and spectral changes in voiceless versus voiced environments may arise from a single mechanism.

For many conditioning environments, the model of articulatory undershoot (Lindblom, 1963) successfully accounts for the links between duration and spectral changes in vowels. This model posits that at short durations, speakers do not have sufficient time to fully achieve articulatory targets, and the resulting hypo-articulation (Lindblom, 1990) manifests itself as centralization of vowel formants. Data from vowels produced at normal versus fast speech rates, and in unstressed versus stressed positions, support this model (Fourakis, 1991; Gay, 1978; Summers, 1987). But data from coda voicing environments contradict it. As Moreton (2004: 2) summarizes, previous production studies have shown that low monophthongal vowels such as [æ] and [ɑ] (i.e., vowels associated with high F1 values) have even higher, not lower, F1 values before voiceless compared to voiced codas, even though duration is shorter. Likewise, high diphthongal offglides, such as that found in /aɪ/ (i.e., offglides associated with low F1 values), have even lower, not higher, F1 values in this environment. For high monopthongal vowels, there is a curious gap in the speech production literature (a few acoustic measurements are reported for the stimuli used in perceptual studies, but these are difficult to interpret systematically:

* Corresponding author.
E-mail addresses: pycha@uwm.edu (A. Pycha), dahan@psych.upenn.edu (D. Dahan).

0095-4470/$ - see front matter © 2016 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2016.01.002

Fischer & Ohde, 1990; Hillenbrand, Ingrisano, Smith, & Flege, 1984; Parker & Walsh, 1981; Walsh & Parker, 1984). We follow the spirit of Moreton (2004) in assuming that the relevant generalization most likely concerns low and high vowel qualities, rather than monophthongal versus diphthongal status per se. If this is the case, the generalization is that formants for both high and low vowel qualities peripheralize before voiceless codas, and centralize before voiced codas. The problem for the undershoot account, then, is that "[i]t seems that pre-voiceless peripheralization occurs *despite* the general effects of vowel shortening, which are antagonistic to it" (Moreton, 2004: 27).

Building upon Thomas (2000), Moreton (2004) speculates that peripheralization occurs because voiceless obstruents are more forceful than voiced obstruents, and the forcefulness of the consonant gesture spreads leftwards onto the preceding vowel, causing peripheralization of formants. Such a spreading analysis has certain advantages: it links spectral qualities to the voicing qualities of the following coda, and it can account for the fact that, in Moreton's data as well as in previous production studies, hyper-articulation does not affect all parts of the vowel equally, but occurs primarily at the vowel offset near a voiceless coda. Yet the analysis also has some serious disadvantages: the concept of "forcefulness" is not fully defined, and it treats vowel durations and spectral qualities as completely separate phenomena.

Beddor (2009) considers a different yet related puzzle, and proposes a solution within the framework of articulatory phonology. In CVNC contexts, American English vowels exhibit longer periods of co-articulatory nasalization before voiceless codas (as in *bent* [bε̃nt]) than before voiced ones (as in *bend* [bε̃nd]) (Cohn, 1990; Raphael, Dorman, Freeman & Tobin, 1975). Beddor (2009) argues that this pattern originates from differences in the relative timing of the tongue body gesture for V and the velum lowering gesture for N, as a function of the voicing of the final C. The gesture for N maintains a roughly equivalent duration across different contexts, but slides leftward or rightward in time relative to the preceding gesture for V. In voiceless contexts, the V and N gestures are close to one another in time. Because the tongue body and velum are independent articulators that can both maintain their individual constrictions simultaneously, this situation creates co-articulatory overlap that manifests itself acoustically in (a) relatively long durations for vowel nasalization and (b) relatively short durations for the nasal consonant itself. In voiced contexts, by contrast, the V and N gestures are separated from one another in time, with acoustic manifestations of (a) relatively short durations for vowel nasalization and (b) relatively long durations for the nasal consonant. Notably, co-articulatory overlap also manifests itself perceptually: American English listeners do better at predicting upcoming codas when they listen to a nasalized vowel, compared to when they listen to a plain vowel (Beddor, McGowan, Boland, Coetzee, & Brasher, 2013). Beddor's (2009) proposal thus draws upon several key tenets of articulatory phonology, namely that individual gestures are crucially organized relative to one another in time, that temporal organization can encode phonological contrasts, and that gestural scores make predictions not just in the kinematic domain, but also in the acoustic and perceptual domains (Browman & Goldstein, 1992; Gafos, 2002; Goldstein & Fowler, 2003).

In this paper, we investigate whether a similar solution can provide a unified account of the temporal and spectral changes that vowels undergo in coda voicing environments. As with Beddor's (2009) proposal, the key idea lies in the relative timing of adjacent gestures before a coda. To illustrate, consider a pair of words such as *bite* /baɪt/ versus *bide* /baɪd/. Because these words contain the diphthong /aɪ/, the speaker must initiate two different tongue-body gestures, namely lowering for /a/ and raising for /ɪ/, before a tongue-tip gesture (i.e., alveolar closure) for either /t/ or /d/. This is schematized in Fig. 1, which depicts gestural timing in dark continuous lines, and its predicted acoustic manifestations in dashed vertical lines. (Note that the gesture for /t/ is depicted as being inherently longer than the gesture for /d/. This does not have major consequences for the analysis we propose, and we assume that difference arises from an aerodynamic voicing constraint (Ohala, 1997)).

Suppose that in voiceless contexts as in *bite* (top panel), the tongue-body gestures for /a/ and /ɪ/ have relatively short inherent durations, as expected, but are also relatively close to one another in time. This is analogous to the temporal configuration for the V and N gestures in *bent,* except that lowering for /a/ and raising for /ɪ/ require different constrictions within the same articulator, so co-articulatory overlap between them is not possible; instead, the temporal proximity between these gestures would presumably give rise to a relatively rapid transition from tongue body lowering for /a/ to tongue-body raising for /ɪ/. As an important consequence, however, the tongue-body gesture for /ɪ/ would be relatively separated in time from the following tongue-tip gesture for the coda /t/.

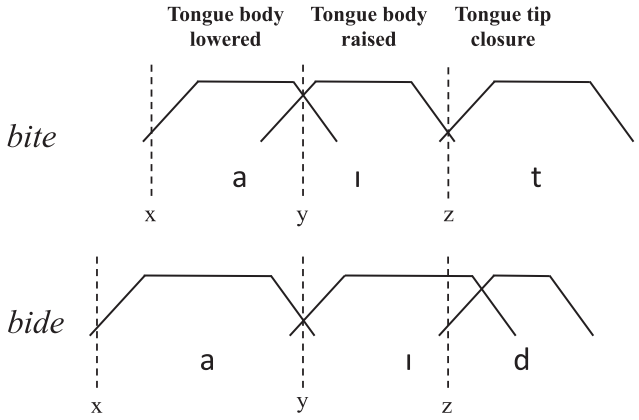


Fig. 1. Proposed gestural scores for words with diphthongal vowels before voiceless codas, as in *bite* [baɪt] (upper panel) and before voiced codas, as in *bide* [baɪd] (lower panel). For clarity, the gestures for the onset consonant are omitted, as is the glottal gesture for the coda. Dashed lines indicate acoustic segmentation.

Note that the tongue tip is semi-independent from the tongue body, so the initiation of the gesture for /t/ could potentially overlap with the preceding gesture for /ɪ/ – although importantly, in voiceless codas, it does not. The gesture for /ɪ/ would therefore have ample time to reach its target.

Furthermore, suppose that in voiced contexts as in *bide* (bottom panel of Fig. 1), the tongue-body gestures for /a/ and /ɪ/ have relatively long inherent durations, as expected, but are also relatively separated from each other in time, just as the gestures for V and N would be in a word like *bend.* Again, since co-articulatory overlap between subsequent gestures made with the same articulator is not possible, the temporal separation between adjacent gestures in the score would presumably give rise to a relatively slow transition from tongue body lowering for /a/ to tongue-body raising for /ɪ/. As an important consequence, however, the tongue-body gesture for /ɪ/ is relatively close to the tongue-tip gesture for /d/. Because these two articulators are semi-independent, the initiation of the gesture for /d/ can overlap with the preceding gesture for /ɪ/ – and importantly, in voiced codas, this is precisely what occurs in the gestural score.

The dashed lines in Fig. 1 show plausible acoustic segmentations that would be produced by these gestural scores: *x* indicates the beginning of [a], *y* indicates the transition from [a] to [ɪ], and *z* shows the end of [ɪ]. We have vertically aligned the top and bottom panels according to *y,* in order to visually orient the reader toward the differences between them. Because co-articulatory overlap cannot occur between /a/ and /ɪ/, *y* is placed equidistant from the end of /a/ and the beginning of /ɪ/, in both top and bottom panels. On the other hand, because co-articulatory overlap can occur between /ɪ/ and the gesture for the coda, in a manner modulated by voicing of the coda, *z* occurs in different locations. In the top panel, with no overlap, it occurs equidistant from the end of /ɪ/ and the beginning of /t/. In the bottom panel, with overlap, it occurs at the beginning of /d/, consistent with the idea that a narrower constriction in the front of the vocal tract (here, the alveolar closure created by the tongue tip) may acoustically mask the effects of a wider constriction further back in the vocal tract (here, the raising of the tongue body) (e.g., Byrd, 1992).

Compared to the *bide* configuration in the bottom panel of Fig. 1, then, the *bite* configuration could conceivably manifest itself in the acoustic record as (a) relatively short overall duration of the diphthong, (b) relatively peripheralized formants for the offglide [ɪ], and (c) a relatively small ratio of [a]-to-[ɪ] duration. Under this scenario, shorter duration and peripheralized formants in words like *bite* (as compared to *bide*) emerge from a single source, and are predicted to occur concomitantly with relatively small duration ratios of [a]-to-[ɪ]. Unlike Moreton's (2004) spreading analysis, our gestural timing proposal does not link spectral changes to the actual voicing qualities of the following coda. But it does link changes in vowel duration to changes in vowel spectral qualities. Furthermore, our gestural timing proposal expands upon a solution already offered for a different problem, namely vowel nasalization in CVNC words, suggesting that it could eventually generalize to other puzzles triggered by coda voicing.

We tested the plausibility of this gestural timing account in one production study and one perception study, both using the American English diphthong /aɪ/. The decision to focus on /aɪ/ was motivated by several factors. First, we wanted to use a diphthong rather than a monophthong, because the proposal of Beddor (2009) involved the relative timing of two individual gestures occurring before a coda, one for V and one for N. Similarly, diphthongs also involve the relative timing of two individual gestures occurring before a coda, e.g., one for the nucleus /a/ and one for the offglide /ɪ/. This similarity allows us to evaluate the generalizability of the gestural timing account, an option that is not as readily available with monophthongs, which can exhibit some spectral movement but arguably involve only a single gesture.

Second, we wanted to focus on /aɪ/ because of previous research focusing on this specific diphthong. For example, Iskarous (2005) collected kinematic data illustrating that /aɪ/ exhibits two distinct constrictions, and that the tongue makes a discrete, not continuous, transition from the first constriction to the second. These results lend plausibility to our conceptualization of /aɪ/ as a sequence of two distinct gestures. In a related study, Iskarous, Nam & Whalen (2010) showed that listeners gave higher naturalness ratings to tokens of /aɪ/ synthesized by a model vocal tract with parameters for discrete transitions, compared to one with continuous transitions. These results suggest that the distinctness between the gestures for /a/ and /ɪ/ manifests itself in the acoustic record, and that we can test gestural hypotheses about /aɪ/ by taking acoustic measurements (see also Krivokapić & Byrd (2012)), who demonstrate that the gestural origins of the acoustic record are available to listeners during perception of phrasal boundaries, and Noiray, Iskarous, and Whalen, (2014), who showed that variability in English vowels is comparable in articulation and acoustics).

In addition, previous production studies report that when /aɪ/ occurs before voiced codas, the acoustic duration of the [a] lengthens more than the acoustic duration of the [ɪ] (a similar result holds for certain other American English diphthongs; Gay, 1968; Jacewicz, Fujimura, & Fox, 2003; Lehiste & Peterson, 1961; Moreton, 2004; Thomas, 2000). This finding is consistent with the gestural scores sketched in Fig. 1, and provides one important reason why we think our hypothesis may be correct. No analogous data is available for monophthongs, whose internal components, if any, are more difficult to separate acoustically than those of diphthongs. Finally, the internal acoustic components of [aɪ] differ significantly from one another, with high F1 and low F2 values in the nucleus [a], but low F1 and high F2 values in the offglide [ɪ]. This makes it relatively easy to pinpoint the transition from the nucleus and offglide in the acoustic record, which is crucial for estimating the degree of temporal proximity or separation between their respective gestures. Pinpointing this transition in other American English diphthongs, such as [eɪ] or [aʊ] (where the latter is typically realized as [aɔ] in most varieties (Thomas, 2001)), while possible, would be less straightforward because the F1 and/or F2 differences between the nucleus and offglide components are not as great as they are with [aɪ].

In the production study, Experiment 1, we took measurements of [aɪ] from the acoustic record. We focused on the ratio of [a]-to-[ɪ] duration, because our gestural timing hypothesis makes specific predictions about this ratio which, unlike the predictions for overall vowel duration and spectral quality, have not previously been tested in the literature. If our gestural timing hypothesis is correct, we expect to see relatively small ratios in voiceless coda environments, compared to relatively large ratios in voiced coda environments; we also expect increasingly large ratios to correlate with increasing centralization of the formants for [ɪ]. In addition to testing these

basic aspects of our hypothesis, the production study also investigated the stability of gestural timing across external conditioning environments. The framework of articulatory phonology purports to capture situations of phonological contrast, such as the coda voicing distinction in *bite* vs. *bide,* in the same way that it captures variation introduced by allophony or by factors such as phrasal position and speech rate (Browman & Goldstein, 1992; Byrd, 2000; Byrd, Kaun, Narayanan, & Saltzman, 2000). Nevertheless, if a given gestural score does help to encode contrast, we would expect it to be relatively stable, even when other sources of variation are present. This expectation is consistent with previous findings demonstrating that segments generally resist the effects of co-articulation (Manuel, 1990; Öhman, 1966), and the effects of other sources of variability like phrase position (Iskarous & Kavitskaya, 2010), just in those instances when these effects would neutralize a contrast. Given this, we predict that changes in gestural timing should consistently occur for coda voicing contexts, even when other factors, such as phrase-final lengthening or changes in speech rate, have their own influence on articulatory dynamics. Furthermore, given that the voicing contrast is a requirement of the phonological grammar, rather than, say, an optional change in speech rate, we would also expect individual speakers to implement changes in gestural timing in a relatively consistent fashion. In this respect, coda voicing should differ from non-contrastive changes such as phrase-final lengthening, where it has been demonstrated that some speakers use lengthening of individual gestures, while others use changes in gestural timing (Byrd, 2000; Edwards, Beckman, & Fletcher, 1991; Krivokapić & Byrd, 2012). To test these predictions, Experiment 1 examined the acoustic signatures of gestural timing before voiceless versus voiced codas, and evaluated the stability of these signatures across speakers and conditioning contexts.

In the perception study, Experiment 2, we manipulated the duration ratio of [a]-to-[ɪ] in spoken words, and examined how this manipulation affected listeners' interpretation of the voicing of the coda. Previous work has already established that English listeners can make perceptual judgments about the voicing of a coda based upon the preceding vowel's duration (Denes, 1955; Raphael, 1972; Raphael, Dorman, & Liberman, 1980) and spectral qualities (Hillenbrand et al., 1984; Jacewicz et al., 2003; Lowenstein & Nittrouer, 2008; Moreton, 2004; Revoile, Pickett, Holden, & Talkin, 1982; Wolf, 1978). If speakers consistently use gestural scores as in Fig. 1, we expect listeners to be sensitive to their properties as well, and we predict that disturbing the key acoustic manifestation of these scores – namely, the duration ratio of [a]-to-[ɪ] – should alter listeners' perceptual judgments, even when the other temporal and spectral cues remain intact (see Fowler, 2006; Krivokapić & Byrd, 2012; Viswanathan, Magnuson, & Fowler, 2014, and references cited therein). For example, previous work shows that in a "tight" versus "tide" perceptual identification task, peripheralized offglides in synthetic [aɪ] tokens facilitated "tight" responses (Moreton, 2004). If these peripheralized offglides in [aɪ] arise as a specific consequence of temporal proximity between the tongue-body gestures for /a/ and /ɪ/, and listeners are sensitive to this gestural origin, then we predict that disturbing this proximity should alter the pattern of perceptual judgments, even when the peripheralized offglides themselves remain intact. Specifically, we predict that manipulating acoustic $[aɪ]_T$ tokens such that they are consistent with temporal *separation* between the gestures for [a] and [ɪ] should interfere with perceiving these tokens as "tight", even when these tokens retain all other acoustic cues to voiceless codas. To test these predictions, Experiment 2 played manipulated tokens to listeners, while collecting forced-choice identification responses ("bite" vs. "bide") and monitoring eye movements to orthographic targets and competitors.

The account we pursue in this paper treats articulatory gestures as theoretical primitives. This is consistent with the central idea of articulatory phonology, namely that we can decompose the actions of the vocal tract into discrete, independent, recombinable units (Goldstein & Fowler, 2003). Conceived of in this coarse-grained way, gestures form the starting point for any phonological hypothesis, just as underlying forms did in other research traditions, which can then be tested by measuring the predicted kinematic, acoustic, or perceptual outputs. Of course, gestures can also be conceived of as finer-grained physical actions, in which case we require a model, such as task dynamics, to show how gestures could actually create and release constrictions in the vocal tract (Saltzman & Kelso, 1987; Saltzman & Munhall, 1989). As should be apparent, the current study largely omits consideration of finer-grained view of gestures in order to focus on a coarser-grained view. That is, we have used gestures to make a theoretical proposal about diphthongs in voiceless versus voiced coda environments, and we are testing its feasibility by measuring predicted outputs in the acoustic record (Experiment 1) and in perceptual responses (Experiment 2). The spirit in which we pursue this hypothesis is captured aptly by Goldstein and Fowler (2003: 170): "If phonological forms are structures of coordinated gestures, as hypothesized in Articulatory Phonology, then gestural analysis should reveal generalizations that are obscured when phonological form is analyzed in some other way".

2. Experiment 1: production

To investigate the predictions of our gestural timing proposal in the acoustic record, we conducted a speech production study that investigated the diphthong /aɪ/ in three contexts that are known to alter vowel durations: before voiceless versus voiced codas, in phrase-medial versus phrase-final positions, and in fast versus normal speech rates. As introduced in Section 1, the acoustic duration of nucleus [a] relative to the acoustic duration of offglide [ɪ] is a manifestation of the hypothesized proximity, or separation, between the gestures /a/ and /ɪ/ in time, and the resulting separation, or overlap with, the between /ɪ/ and the coda gesture. We therefore focused the bulk of our investigation on the ratio [nucleus duration/offglide duration]. We predicted that this ratio will be smaller before voiceless codas than before voiced codas, and that this difference will hold across speakers and across the conditioning contexts of phrasal position and speech rate. We also predicted that this ratio should correlate, inversely, with offglide qualities: the higher the ratio, the more akin the gestures are to the hypothesized scores in the bottom panel of Fig. 1 (*bide*), and the more centralized the offglide [ɪ] should be.

Although phrasal positions and speech rate are not the primary focus of this study, the literature nevertheless suggests some predictions. Previous work on phrasal lengthening reports inter-speaker variability and therefore suggests that some speakers will exhibit differences in duration ratio across phrase-medial and phrase-final positions, while others will not (e.g., Krivokapić & Byrd 2012). Previous work on speech rate, focused on /aɪ/ as well as other diphthongs, suggests that the duration ratio may change as a function of rate: for example, Gay (1968: Table I) reports duration measurements consistent with the idea that the ratio is smaller at faster rates and larger at slower rates (see also Gottfried, Miller, and Meyer (1993)). Gay (1968) also reports that the rate of F2 change during the offglide remains relatively constant across normal and fast speech, and suggests that the overall trajectory of the offglide gesture remains constant across rates, but becomes clipped – and therefore does not fully achieve its target – in fast speech.

In addition to the ratio and its related value of overall diphthong duration, we also report acoustic measurements for nucleus quality and offglide quality, as indicated by F1 and F2 values. This allows us to confirm that the nucleus and offglide qualities are affected according to gestural timing in the manner that we proposed in Section 1.

2.1. Method

2.1.1. Stimuli

Stimuli were short sentences that crossed two factors: coda voicing and phrasal position, as depicted in Table 1. During production, these were crossed with a third factor, speech rate, which was either “normal” or “faster”.

Six pairs of words served as targets: *bite*∼*bide*, *height*∼*hide*, *sight*∼*side*, *tight*∼*tide*, *write*∼*ride*, and *white*∼*wide.* (Although some American English speakers produce the initial consonant of *white* as [ʍ], all of the participants in Experiment 1 produced [w] for both *white* and *wide).* Each pair consisted of CVC monosyllables with the diphthong /aɪ/ before a voiced or a voiceless coda. Words without a coda, such as *buy,* were also included in the stimulus set but their data is not reported here. Participants produced each sentence at two different rates, normal and faster, determined individually by each participant after instruction by the experimenter.

The target words differ in their semantic and syntactic characteristics. In order to maintain controlled conditions despite these differences, short context-neutral sentences were used. Participants were asked to imagine a listener seated in front of a computer screen displaying the words *bite* and *bide.* They were told to pronounce each sentence as if they were instructing the listener to move one of the words leftwards.

2.1.2. Participants

Nine individuals, all female native speakers of American English between the ages of 18 and 30, participated. Three participants were from Philadelphia, one was from South Central Pennsylvania, one was from Columbus, Ohio, one was from Long Island, New York, one was from Seattle, Washington, and two were from Southern California. All were living in Philadelphia at the time of the recording.

2.1.3. Procedure

The stimulus sentences were randomized, and presented to participants as a printed list. Beginning with the normal speech rate, the participants read each sentence three times before proceeding to the next sentence on the list. Participants then repeated this procedure with the fast speech rate. All recordings took place in a quiet room using a high-quality head-mounted microphone connected to a laptop computer sampling at 44,100 Hz. An experimenter monitored each recording session. When errors in fluency or pronunciation occurred, the experimenter asked the participant to repeat the sentence.

2.1.4. Acoustic analysis

The diphthongs were segmented using waveforms and spectrograms in Praat (Boersma & Weenink, 2012), according to the following procedure. For words beginning with an obstruent consonant or [h], the diphthong began at the onset of periodic voicing. For words beginning with [ɹ], the diphthong began at the point where F3 started to rise, while for words beginning with [w], the diphthong began at the point where both F1 and F2 started to rise and separate from one another. For all words, the diphthong ended when the spectrogram showed no more evidence for the presence of F2. Participants sometimes produced creaky voice, particularly in phrase-final positions; creaky portions were included as long as glottal pulses were still evident.

Four measurements were taken for each diphthong: nucleus duration and offglide duration, which were used to calculate the total duration and duration ratio, as well as nucleus quality and offglide quality. The duration measurements are depicted in Fig. 2 (left panel). Nucleus duration was calculated from the beginning of the diphthong as determined by the segmentation procedure, to the point of maximum F1. This point was chosen because it served as a simple yet reliable marker of the end of the nucleus (where the formants approximate a steady state), and the beginning of the offglide (where the formants are in rapid transition). Offglide duration was calculated from the point of maximum F1, to the end of the diphthong as determined by the segmentation procedure. The

Table 1
Stimulus sentences for Experiment 1.

Coda	Medial	Final
Voiceless	Move bite left.	Move bite. Left.
Voiced	Move bide left.	Move bide. Left.

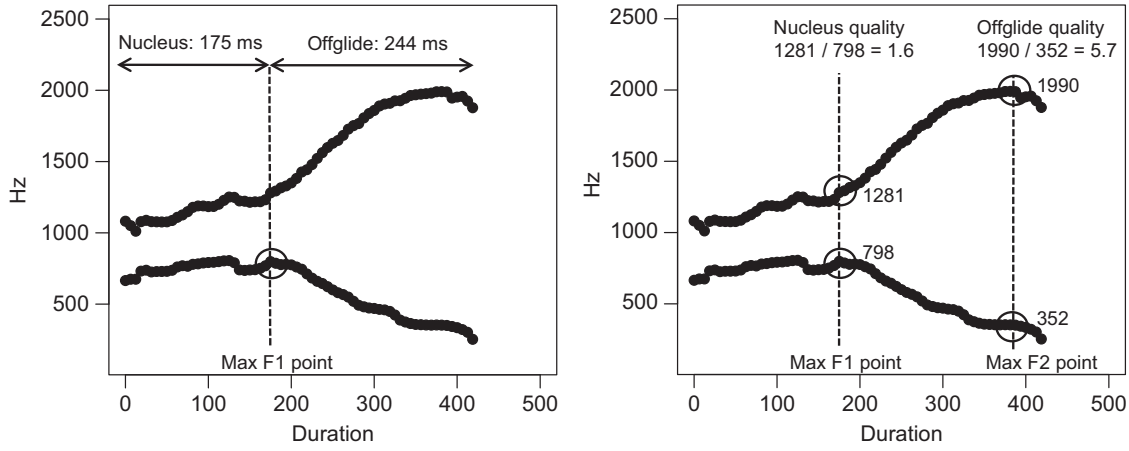


Fig. 2. Sample duration (left panel) and vowel quality (right panel) measurements for Experiment 1. The x-axis shows duration in milliseconds, and the *y*-axis shows frequencies in Hertz. The formant trajectories were plotted from a Praat formant object, creating using the "Burg" method of linear predictive coding with 10 coefficients (i.e., maximum of five formants) and a maximum formant value of 5500 Hz.

nucleus and offglide durations were used to calculate our variables of interest, total duration and duration ratio. Thus, for the diphthong in the Fig. 2, the total duration = 175 + 244 = 419 ms, and the duration ratio = 175 ms/244 ms = 0.72.

The quality measurements are depicted in Fig. 2 (right panel). The nucleus formants were measured at the time point where F1 reached its maximum value. This time point represents the target low articulation for /a/ and also replicates the procedure used in Moreton (2004); note, however, that the F2 value at this time point does not necessarily represent the target back articulation for /a/. The offglide formants were measured at the point where F2 reached its maximum value. This time point represents the target front articulation for /ɪ/ and replicates the procedure used in Moreton (2004); note, however, that the F1 value at this time point does not necessarily represent the target high articulation for /ɪ/. After the nucleus and offglide measurement points were located, the qualities of nucleus and offglide were then calculated as the ratio of F2/F1 at that point. Smaller values of this ratio mean that F1 and F2 are close together, while larger values indicate that F1 and F2 are further apart. For nuclei, then, smaller values indicate peripheralization toward [a] while larger values indicate centralization toward [ʌ]. For offglides, on the other hand, smaller values indicate centralization toward [ɪ] while larger values indicate peripheralization toward [i].

All measurements were produced automatically with a Praat script. First, we used each participant's audio recording to create a Praat formant object using the "Burg" method. As the Praat documentation states, this process uses linear predictive coding to approximate the spectrum of each analysis frame by a number of formants. We used the following settings: time-step = 0, maximum number of formants = 5, maximum formant = 5500 Hz, window length = 0.025, pre-emphasis = 50 dB. However, for participants 1, 2, 5, and 9, the maximum formant was adjusted to 6200 Hz in order to eliminate spurious formants that occurred at lower settings. Second, we ran a script that searched the formant object for the timepoints of maximum F1 value and maximum F2 values within each diphthong, using parabolic interpretation. As the Praat documentation states, this method is most appropriate for instances in which *x* is a smooth function of *y*, as in formant tracking. Once those timepoints were located, the script measured both F1 and F2 at those points, and also measured duration from onset to maximum F1 (nucleus duration), and duration from maximum F1 to offset (offglide duration).

Note that Praat estimates formant values for a particular time point by integrating information from a 50 ms time window centered on that point. In order to accurately calculate the minimum and maximum F1 values, particularly when these occurred very close to the beginning or end of the vowel, the script examined formants starting 25 ms prior to the segmented beginning of each vowel, and ending 25 ms after its segmented ending. Note also that the linear predictive coding method used to create the Praat formant objects shows some attraction to harmonics; this can be seen in Fig. 2, where the formant tracks display some sudden movements, even though the formants themselves probably did not exhibit such movements. Approximately one-fourth of the tokens were visually examined by the experimenters in order to ensure the accuracy of the maximum F1 time points; no significant discrepancies were found.

Since each participant produced three repetitions of each stimulus at two different rates, this yielded 1296 tokens of /aɪ/ for analysis (6 word pairs × 2 coda voicing values × 2 phrase positions × 2 speech rates × 3 repetitions × 9 participants). With this data set, we examined each outcome variable with a linear mixed-effects model, as implemented by the lme() function in the nmle package in R. Duration ratio, nucleus quality, and offglide quality are all ratio values and therefore potentially bounded below by 0. To make these outcome variables suitable for analysis, we transformed them with logarithms, although note that for all tables and figures, we have retained the original, non-transformed values. The predictor variables for each analysis were Coda voicing (baseline value = voiceless), Phrasal position (baseline value = medial), and Speech rate (baseline value = fast). Both participant and repetition were entered as random factors. The complete results of all four models are displayed in Appendix A.

2.2. Results: duration and duration ratio

Table 2 summarizes the descriptive results for duration and duration ratio.

Table 2
Results from Experiment 1. Means and standard deviations for overall duration (in milliseconds) and duration ratio [nucleus duration/offglide duration] of the diphthong [aɪ] across three different conditioning environments.

	Coda voicing		Phrasal position		Speech rate	
	Voiceless	Voiced	Medial	Final	Fast	Normal
Overall duration	151.6	191.5	149.7	193.5	137.0	206.2
	(48.9)	(69.7)	(55.6)	(63.3)	(47.8)	(58.1)
Duration ratio	0.25	0.38	0.32	0.32	0.37	0.30
	(0.22)	(0.23)	(0.25)	(0.22)	(0.25)	(0.22)

Overall durations were longer in voiced coda contexts compared to voiceless, although this difference only approached significance ($\beta=6.46$, $t=1.86$, $p=0.06$). In addition, overall durations were longer in phrase-final positions compared to phrase-medial ($\beta=38.94$, $t=11.21$, $p<0.05$), and at normal speech rates compared to fast rates ($\beta=55.55$, $t=15.99$, $p<0.05$). Despite the basic similarity in the way these three conditioning contexts affected the duration of the entire vowel, their effects on the duration ratio were very different. Duration ratio was significantly greater in voiced coda contexts compared to voiceless contexts ($\beta=0.33$, $t=2.56$, $p<0.05$), did not change in phrase-final positions compared to phrase-medial positions ($\beta=0.05$, $t=0.39$, $p=0.70$), and was significantly smaller at normal speech rates compared to fast rates ($\beta=-0.43$, $t=-3.33$, $p<0.05$).

Fig. 3 shows the trends and spread in our data, coded according to voicing (a), phrasal position (b), and speech rate (c). Overall, we see that the duration of [ɪ] spans a wider range of values than duration of [a]. Also, while both components of the diphthong tend to increase as total duration increases, the duration of [ɪ] is more tightly correlated to it (Pearson's $r=0.91$, $t=77.77$, $p<0.05$), compared to duration of [a] (Pearson's $r=0.62$, $t=28.45$, $p<0.05$). In Fig. 3a (left), the solid regression line for voiced environments lies well above the dashed line for voiceless ones, at least at longer durations, which impressionistically suggests that for any given duration of [aɪ], the corresponding duration of nucleus [a] will tend to be greater in a voiced environment than in a voiceless one. We see the opposite pattern in Fig. 3a (right), which suggests that the duration of offglide [ɪ] will tend to be smaller in a voiced environment than in a voiceless one. Thus, the significant differences in duration ratio between voicing environments would appear to arise from changes to both diphthong components simultaneously: roughly speaking, in voiced environments, nuclei get longer and offglides get shorter.

In Fig. 3b, in both panels, the solid line for phrase-final environments is nearly identical to the dashed line for phrase-medial environments, which suggests that the equivalence in duration ratio between phrasal environments derives from an actual equivalence in diphthong components (and not, say, from opposing changes to individual components that could potentially cancel each other out in a ratio calculation). In Fig. 3c (left), the solid line for normal speech rates lies somewhat below the dashed line for fast rates, at least at lower durations, while we see the opposite pattern in Fig. 3c (right). This suggests that the modest yet significant differences in duration ratio between speech rates arise from changes to both diphthong components simultaneously: roughly speaking, in normal speech rates, nuclei get shorter and offglides get longer, compared to those in fast speech rates.

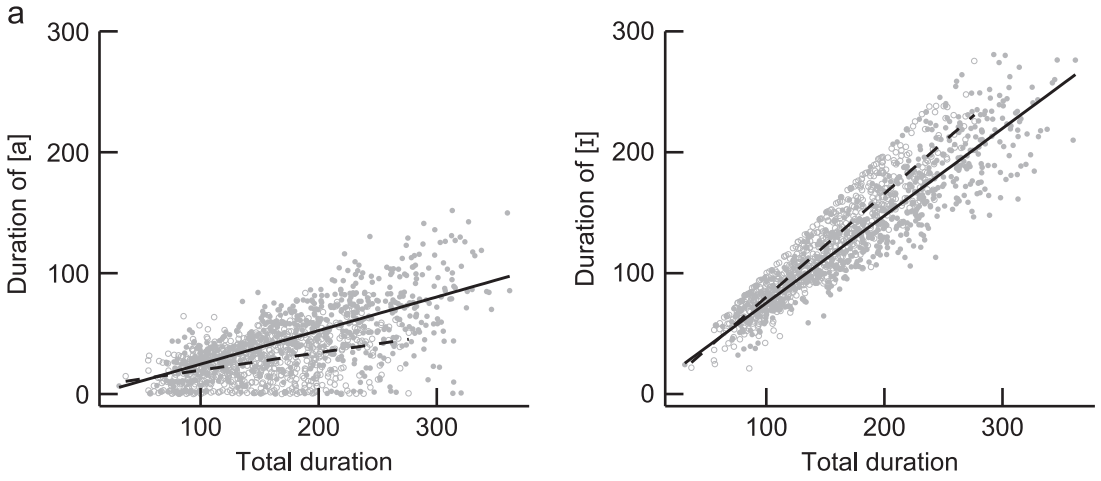
2.3. Stability of duration ratio across contexts

In order to assess the stability of changes in duration ratio associated with coda voicing, we examined the effect of voicing for each of the four contexts in which it was tested and found it to be consistent across contexts (see Table 3), with smaller duration ratios before voiceless consonants than before voiced ones. At fast speech rates in medial positions, the mean ratio was 0.30 versus 0.37 in voiceless versus voiced coda contexts, a difference of 0.07; in final positions, it was 0.28 versus 0.39, a difference of 0.11. At normal speech rates in medial positions, the mean ratio was 0.22 versus 0.39 in voiceless versus voiced coda contexts, a difference of 0.17; in final positions, it was 0.22 versus 0.38, a difference of 0.16. As these values indicate, the change in ratios between voiceless and voiced contexts was smaller at fast rate than at normal rate ($\beta=0.3$, $t=2.88$, $p<0.05$), probably as a result of the compression of the total vowel duration at fast rate compared to normal rate. No other interactions between voicing, phrasal position, and speech rate were significant.

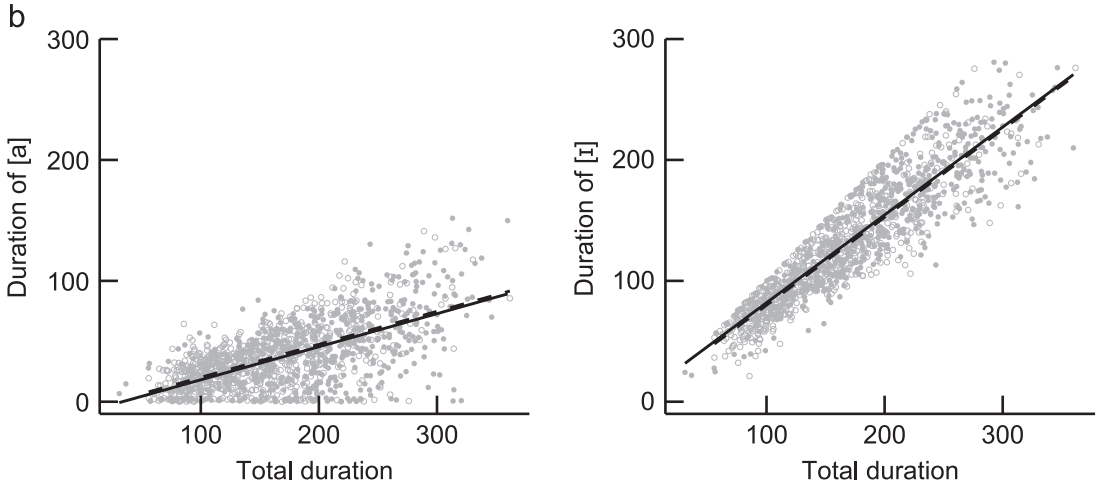
2.4. Stability of duration ratio across speakers

Coda voicing also exerted consistent effects across speakers. As Table 4 (left columns) shows, duration ratio was greater in voiced coda contexts (compared to voiceless) for all nine participants.

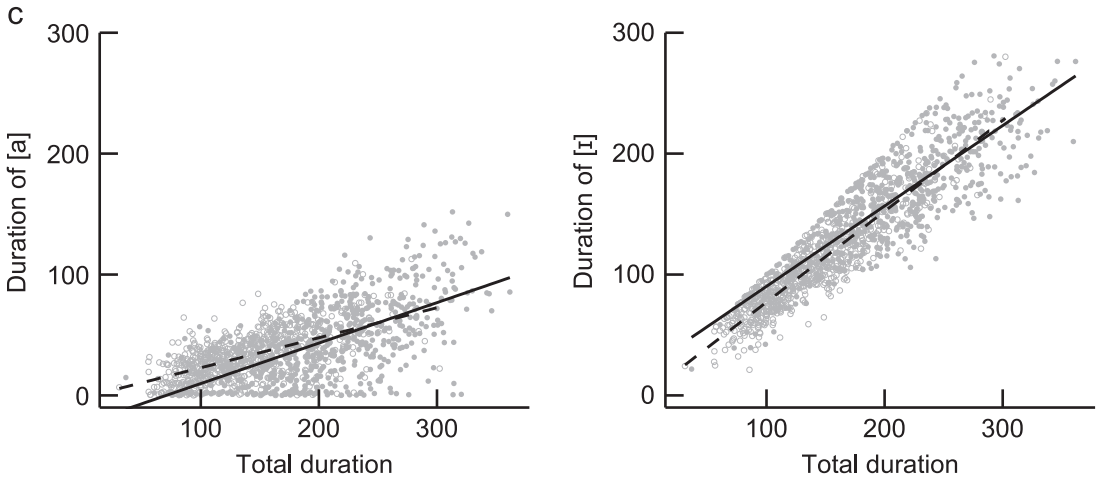
Phrasal position, on the other hand, exerted different effects for different speakers. As Table 4 (middle columns) shows, duration ratio was greater in final position (compared to medial) for participants 1 and 3, roughly equivalent across positions for participants 2, 4, and 8 and smaller in final position (compared to medial position) for participants 5, 6, 7, and 9. Speech rate showed relatively consistent effects across speakers. As Table 4 (right columns) shows, duration ratio was smaller or equivalent at normal speech rates (compared to faster) for all nine participants.



Voicing: Open circles and dashed line indicate voiceless environments, filled circles and solid line indicate voiced environments.



Phrasal position: Open circles and dashed line indicate phrase-medial environments, filled circles and solid line indicate phrase-final environments.



Rate: Open circles and dashed line indicate fast speech rate, filled circles and solid line indicate normal speech rate.

Fig. 3. (a, b, c): Scatterplots depicting relationship between total duration of the diphthong and nucleus [a] (left panels), and the offglide [ɪ] (right panels).

Table 3
Means and standard deviations for the acoustic measurements in Experiment 1.

		Fast		Normal	
		Medial	Final	Medial	Final
Duration (ms)	Voiceless	107.9 (26.3)	146.8 (41.4)	163.4 (33.9)	188.4 (51.1)
	Voiced	114.3 (32.6)	179.0 (49.7)	213.2 (47.5)	259.7 (49.5)
Duration ratio	Voiceless	0.30 (0.30)	0.28 (0.18)	0.22 (0.17)	0.22 (0.20)
	Voiced	0.37 (0.24)	0.39 (0.22)	0.39 (0.23)	0.38 (0.23)
Nucleus (F2/F1)	Voiceless	2.5 (0.3)	2.3 (0.3)	2.2 (0.3)	2.1 (0.3)
	Voiced	2.2 (0.3)	2.0 (0.3)	1.8 (0.2)	1.8 (0.2)
Offglide (F2/F1)	Voiceless	4.3 (0.8)	5.2 (0.9)	5.5 (0.9)	6.0 (0.9)
	Voiced	4.0 (0.7)	4.6 (0.7)	4.7 (0.7)	5.1 (0.8)

Table 4
Means and standard deviations for duration ratios [nucleus duration/offglide duration] in diphthong [aɪ] across three environments, for nine individual participants.

Participant	Coda voicing		Phrasal position		Speech rate	
	Voiceless	Voiced	Medial	Final	Fast	Normal
1	0.22	0.39	0.22	0.39	0.30	0.30
	(0.25)	(0.29)	(0.19)	(0.33)	(0.27)	(0.30)
2	0.24	0.42	0.33	0.33	0.36	0.31
	(0.15)	(0.21)	(0.22)	(0.20)	(0.19)	(0.22)
3	0.18	0.34	0.25	0.27	0.28	0.25
	(0.11)	(0.16)	(0.13)	(0.18)	(0.16)	(0.16)
4	0.39	0.46	0.43	0.42	0.45	0.40
	(0.22)	(0.22)	(0.27)	(0.17)	(0.24)	(0.21)
5	0.15	0.26	0.23	0.18	0.21	0.20
	(0.11)	(0.19)	(0.18)	(0.14)	(0.15)	(0.18)
6	0.28	0.34	0.32	0.30	0.32	0.30
	(0.26)	(0.28)	(0.26)	(0.28)	(0.30)	(0.24)
7	0.27	0.47	0.41	0.33	0.38	0.36
	(0.14)	(0.19)	(0.22)	(0.17)	(0.18)	(0.22)
8	0.28	0.35	0.32	0.31	0.33	0.30
	(0.37)	(0.20)	(0.39)	(0.17)	(0.37)	(0.20)
9	0.27	0.42	0.36	0.33	0.38	0.31
	(0.18)	(0.22)	(0.22)	(0.22)	(0.22)	(0.20)

2.5. Results: nucleus and offglide quality

Results of linear models showed that nuclei were more peripheralized before voiced codas (i.e., smaller F2/F1 ratios for [a]) compared to voiceless ones ($\beta = -0.10$, $t = -8.51$, $p < 0.05$). This could possibly reflect the influence of Canadian raising alternations, although the nucleus was also more peripheralized in final phrasal position than in medial position ($\beta = -0.06$, $t = -5.04$, $p < 0.05$), and in normal speech rate compared to fast ($\beta = -0.13$, $t = -10.65$, $p < 0.05$).

Results also showed that offglides were more centralized before voiced codas (i.e., smaller F2/F1 ratios for [ɪ]) than before voiceless codas (($\beta = -0.07$, $t = -5.34$, $p < 0.05$). These effects of voicing were consistent across all the sub-conditions (in fast speech and medial position, in fast speech and final position, in normal speech and medial position, and in normal speech and final position, see Table 3). As expected, the effects of phrasal position and speech rate on nucleus and offglide qualities were different from those of voicing. The offglide was more peripheralized (i.e., larger F2/F1 ratios for [ɪ]) in final position compared to medial ($\beta = 0.19$, $t = 14.38$, $p < 0.05$) and in normal speech compared to fast ($\beta = 0.25$, $t = 18.68$, $p < 0.05$).

2.6. Relationship between duration ratio and offglide quality

Across all tokens, there was a modest but significant negative correlation between duration ratio and offglide quality (Pearson's $r = -0.38$, $t = -15.00$, $p < 0.05$), depicted in Fig. 4.

2.7. Discussion

Overall, results of Experiment 1 confirm acoustic predictions made by the gestural scores in Fig. 1. The ratio [nucleus duration/offglide duration] for the diphthong [aɪ] was significantly smaller in voiceless coda contexts than in voiced coda contexts, a finding that is also consistent with previous reports. The pattern of smaller ratios in voiceless versus voiced contexts occurred in every sub-condition of the experiment: even when phrasal position and speech rates exerted their own effects on the diphthong [aɪ], coda voicing context still governed differences in duration ratio. Furthermore, the pattern of smaller ratios in voiceless versus voiced contexts was exhibited by all nine of the individual speakers who participated in the experiment. This is an important point because

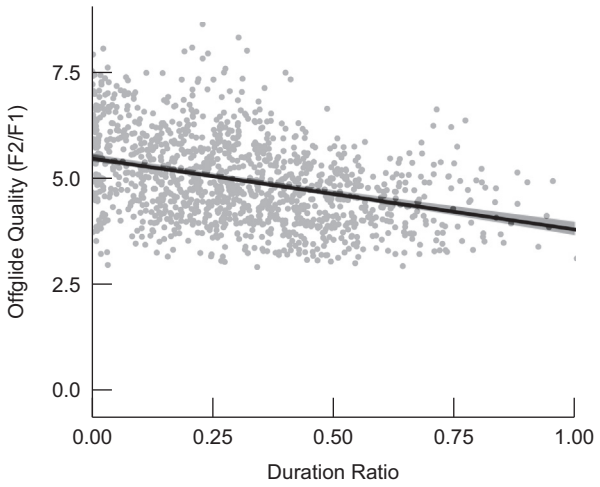


Fig. 4. Scatterplot depicting the inverse relationship between duration ratio and offglide quality (F2/F1) in Experiment 1.

previous investigations into gestural timing in other contexts have reported substantial inter-speaker variability. By contrast, the inter-speaker consistency in our data support the idea that different gestural timing relationships before voiceless versus voiced codas contribute to the creation of phonological contrast.

The spectral data show that nuclei were centralized and offglides were peripheralized in voiceless coda contexts, compared to voiced coda contexts. This finding is consistent with previous reports, and also with our proposed gestural scores. In addition, increased values for [nucleus duration/offglide duration] correlated with decreased values for offglide quality (i.e., smaller F2/F1 ratios for [ɪ], indicating centralization), lending support to our idea that offglide quality derives specifically from temporal relationships between the gestures for /a/ and /ɪ/.

Although not the primary focus of Experiment 1, the data for phrasal lengthening nevertheless suggest a few observations. While overall diphthong durations were significantly larger in final compared to medial positions, the overall ratios of [nucleus duration/offglide duration] did not differ. Inspection of individual participant data for duration ratio shows inter-speaker variability, with some speakers exhibiting an increase across positions, others a decrease, and others no change at all. This variability is consistent with previous kinematic studies demonstrating that some speakers employ temporal realignment of gestures in phrase-final positions while others do not. The consistency between our findings for duration ratio and kinematic studies adds some credence to the use of duration ratio as a proxy for the temporal dynamics of the articulatory gestures associated with the production of diphthong components.

An open question concerns the results for speech rate. Results from Experiment 1 indicate that the duration ratio was larger at fast rates than at normal rates. This is inconsistent with Gay (1968: Table I), who reports smaller duration ratios at fast rates, and larger ratios at normal rates, although this may be due to differences in segmentation criteria. Nevertheless, our results are consistent with Gay's basic proposal that the /ɪ/ gesture undergoes temporal clipping at fast rates, and our vowel-quality analysis bolsters this idea. If clipping occurs without any temporal changes to the /a/ gesture and without changes in timing between /a/ and /ɪ/ gestures, the outcome would be larger duration ratios at fast rates compared to normal rates, such as we report here.

In sum, results of Experiment 1 support the notion that the duration and spectral properties of vowels before voiceless versus voiced codas in American English originate in changes to gestural timing.

3. Experiment 2: perception

Experiment 2 tests the predictions of our gestural timing proposal in the perception domain, asking whether listeners are sensitive to the hypothesized proximity versus separation of gestures when they make coda voicing judgments of words containing [aɪ]. Like Experiment 1, Experiment 2 focuses on the ratio [nucleus duration/offglide duration]. We created two sets of spoken words: congruent stimuli, in which duration ratios of [aɪ] were appropriate to the coda context (shorter ratios in voiceless contexts, $[aɪ]_{T\text{-}Short}$, than in voiced contexts, $[aɪ]_{D\text{-}Long}$), and incongruent stimuli, in which duration ratios of [aɪ] were inappropriate to the coda context (longer ratios in voiceless contexts, $[aɪ]_{T\text{-}Long}$, than in in voiced contexts, $[aɪ]_{D\text{-}Short}$). Importantly, we did not manipulate any other variables, so the total diphthong duration, as well as the nucleus and offglide qualities, remained contextually appropriate in all stimuli; codas themselves also remained intact. Thus, incongruent stimuli exhibited a mismatch between the relative durations of [a] and [ɪ] and the rest of the diphthong's properties – that is, individual spectral cues to coda voicing were maintained in the diphthong, but their gestural basis was not. We presented these stimuli to listeners in a forced-choice identification task: for example, after hearing $[baɪt]_{T\text{-}Long}$, participants clicked on a word indicating whether they heard “bite” or “bide”. We tracked eye movements as they completed the task. If listeners are sensitive to the gestural origins of the temporal and spectral qualities of [aɪ], we expect to see a difference in participant interpretations of congruent versus incongruent stimuli, manifested in their ultimate choice and/or in the

degree to which they fixate (i.e., consider) each word option before making their choice. Specifically, we expect more errors in mouse-click responses, and/or more time spent looking at the competitor word, when the duration ratio is incongruent with the other voicing cues (i.e., [baɪt]$_{\text{T-Long}}$ and [baɪd]$_{\text{D-Short}}$) compared to when it is congruent (i.e., [baɪt]$_{\text{T-Short}}$ and [baɪd]$_{\text{D-Long}}$).

Moreton's (2004) perceptual experiment showed that shorter values for nucleus duration facilitated "tight" responses in a "tight" versus "tide" identification task. However, Moreton's manipulation of nucleus duration differs fundamentally from the manipulation that we use in Experiment 2. In his stimuli, nucleus duration and overall diphthong duration always changed in tandem; for example, a 14 ms decrease in nucleus duration created a 14 ms decrease in overall duration. This makes it impossible to separate the effects of nucleus duration from those of overall vowel duration, which have already been shown to influence perception of coda voicing (Denes, 1955; Raphael, 1972; Raphael et al., 1980). In the incongruent stimuli that we created, by contrast, nucleus duration changed independently of overall duration; for example, a 14 ms decrease in nucleus duration created no change in overall duration (instead, it created a 14 ms increase in offglide duration).

We used an eye-tracking methodology to assess listeners' perceptual interpretation of the stimulus words. When eye-tracking is employed to investigate spoken-language comprehension, participants are typically presented with a small set of visual images or printed words on the computer screen. At the same time, they hear a spoken stimulus. The participant's task is to click on the image or word that represents what they heard (e.g., Allopenna, Magnuson, & Tanenhaus, 1998). For example, after hearing a word such as *beaker,* the participant clicks on the image of a beaker. As the participant completes the task, a camera collects his or her eye movements, which reveal where the participant's gaze falls on the screen. This allows the experimenter to monitor which visual images or printed words the participant looks at, beginning with the onset of the spoken stimulus and ending when he or she settles on an interpretation by clicking the mouse.

The assumption is that the participants' eye fixations (i.e., moments when the eyes are still, and fixated on a particular object) indicate how they interpret the spoken word as the phonetic information from the spoken stimulus unfolds. For instance, Allopenna et al. (1998) showed that, as participants heard the early portion of a word, e.g., during the first sounds of *beaker*, they tended to look at the picture of a beetle more than on that of a carriage. As the word continued to unfold, however, participants' fixations started converging on the picture of the beaker, and away from that of the beetle. Importantly, these fixations took place before participants made their final decision by clicking on the target image, e.g., the image of the beaker. Thus, participants' fixations over time can reveal their evolving interpretation of a spoken word before they make an overt response. Subsequent research has shown that people's fixations are sensitive to the presence of brief phonetic attributes, such co-articulatory information in a vowel associated with the following consonant (e.g., Dahan, Magnuson, & Tanenhaus, 2001; Beddor et al., 2013; Salverda, Kleinschmidt, & Tanenhaus, 2014). These findings suggest that eye-tracking may be a particularly apt methodology for revealing the perceptual impact of relatively subtle and brief manipulations of vowels.

3.1. Method

3.1.1. Stimuli

Targets consisted of fourteen pairs of monosyllabic words which contained [aɪ] and contrasted in the voicing of the coda consonant: *bite*∼*bide, bright*∼*bride, dice*∼*dies, fright*∼*fried, lice*∼*lies, life*∼*live, rice*∼*rise, right*∼*ride, sight*∼*side, strife*∼*strive, tripe*∼*tribe, vice*∼*vise, white*∼*wide, write*∼*ride*. The duration ratio of each diphthong was manipulated so that it was either congruent or incongruent with its voicing context, producing a design that crossed two factors: Coda voicing (voiceless versus voiced), and Diphthong duration ratio (congruent versus incongruent), as depicted in Table 5.

Stimulus creation proceeded in several steps. To create bases for the stimuli, a male speaker of American English produced four repetitions of each word at a normal speech rate (in isolation, which makes their durations and other variables somewhat different from those reported in Experiment 1, where speakers produced phrases). We calculated the duration ratio for each repetition using the same procedure as in Experiment 1, as described in Section 2.1.4. From these repetitions, we selected the tokens with the most similar values of the ratio [nucleus duration/offglide duration] to serve as bases. For example, the four repetitions of *bite* had duration ratios of 0.20, 0.28, 0.33, and 0.41. Meanwhile, the four repetitions of *bide* had ratios of 0.43, 0.49, 0.59, and 0.64. Thus, the tokens *bite*-0.41 and *bide*-0.43 were selected as bases because their respective values most closely approached the range of values exhibited by the opposite coda voicing category.

Next, we established target values by selecting the tokens with the most dissimilar duration ratios. Thus, 0.20 was selected as the *bite* target value because it was the smallest for its voicing category (the most "bite"-like), while 0.64 was selected as the *bide* target value because it was the largest for its category (the most "bide"-like).

To create stimuli for the congruent condition, each base token was manipulated such that its duration ratio matched the target value of its voicing category. For example, the *bite*-0.41 base token was manipulated so that its duration ratio became 0.21. This base

Table 5
Stimulus design for Experiment 2.

	Voiceless	Voiced
Congruent ratio	[baɪt]$_{\text{T-Short}}$	[baɪd]$_{\text{D-Long}}$
Incongruent ratio	[baɪt]$_{\text{T-Long}}$	[baɪd]$_{\text{D-Short}}$

token originally had an overall duration of 224 ms, with a nucleus of 65 ms and an offglide of 159 ms. Using the Praat duration manipulation module, the nucleus portion was decreased to 38 ms and its offglide was increased to 186 ms, achieving the target duration ratio of 0.21 (=39 ms/185 ms) (for details, see Appendix B). The overall duration of the base diphthong remained unchanged, as did its nucleus and offglide qualities. This is depicted in Fig. 5, top panel, where dashed lines represent the formant values of the base, and solid lines represent those of the newly-created target.

The *bide*-0.43 base token was manipulated similarly. This base token originally had an overall duration of 469 ms, with a nucleus of 141 ms and an offglide of 328 ms. The nucleus portion was increased to 183 ms and its offglide was decreased to 286 ms, achieving the target duration ratio of 0.64 (=183 ms/286 ms). Again, the overall duration of the base diphthong remained unchanged, as did its nucleus and offglide qualities. This is depicted in Fig. 5, third panel.

To create stimuli for the incongruent condition, a similar process was used, except that now base tokens were manipulated to match the target value of the *opposite* coda voicing category. Thus, the *bite*-0.41 base token was manipulated so that its duration ratio became 0.64 (the *bide* target value), and the *bide*-0.43 base token was manipulated so that its duration ratio became 0.20 (the *bite* target value). Again, the overall duration, nucleus quality, and offglide quality of the base diphthong remained unchanged. This is depicted in Fig. 5, second and fourth panels. Note that all duration manipulation occurred *in situ;* the diphthong /aɪ/ was not spliced from its base word.

This process was repeated for each of the fourteen word pairs, with target values defined individually for each pair on the basis of the speaker's four repetitions. The resulting stimulus set contained 56 test tokens (14 pairs × 2 Voicing conditions × 2 Duration ratio conditions=56), whose overall acoustic characteristics are displayed in Table 6. Importantly, the procedure used for stimulus creation produced substantial differences between the congruent and incongruent condition for duration ratio, but negligible differences for overall duration, nucleus quality, and offglide quality.

Four lists of test tokens were prepared, with the goal of balancing the presentation of token types across participants. For a given list, each of the fourteen word pairs was assigned to group A, B, C, or D, with an approximately equal number of pairs assigned to each group, and with the four lists differing according to the pair/group assignment. Each list consisted of four ordered blocks. In

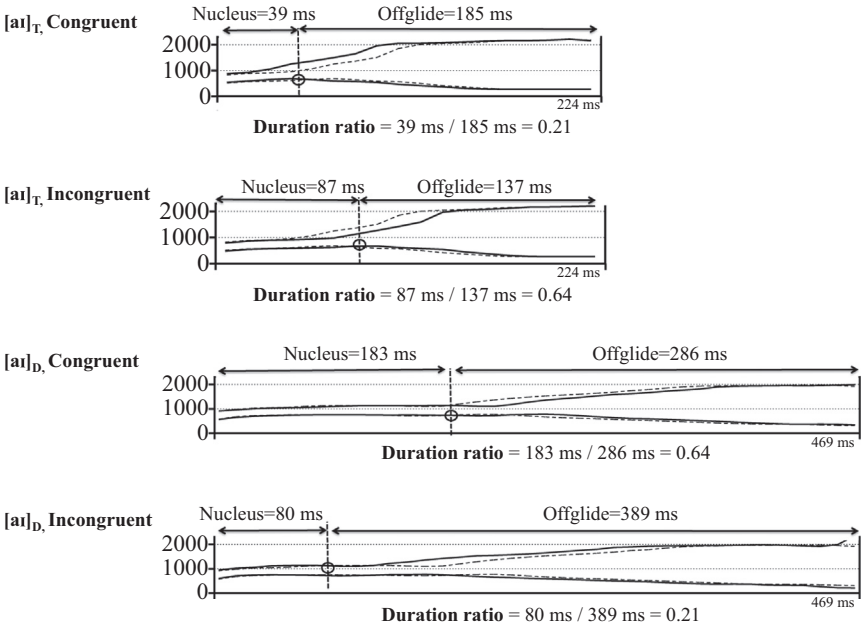


Fig. 5. Stimulus construction for Experiment 2. The $[aɪ]_T$ vowels occur in voiceless coda contexts (*bite*), while the $[aɪ]_D$ vowels occur in voiced coda contexts (*bide*). Congruent ratios are contextually appropriate (shorter for $[aɪ]_T$, longer for $[aɪ]_D$) while incongruent ratios are contextually inappropriate (longer for $[aɪ]_T$, shorter for $[aɪ]_D$). On each formant track, the *x*-axis represents time in milliseconds, the *y*-axis represents frequencies in Hertz. Dashed lines represent formant values of the base token of /aɪ/. Solid lines represent formant values of the newly-created target token of /aɪ/. Figures drawn to approximate, not exact, scale.

Table 6
Mean acoustic measurements and standard deviations of the diphthongs used in the stimulus tokens for Experiment 2.

	Vowel	Congruent	Incongruent
Duration ratio	$[aɪ]_T$	0.27 (0.07)	0.74 (0.19)
	$[aɪ]_D$	0.77 (0.21)	0.31 (0.21)
Duration (ms)	$[aɪ]_T$	235.9 (27.8)	235.9 (27.8)
	$[aɪ]_D$	466.6 (63.1)	466.6 (63.1)
Nucleus (F2/F1)	$[aɪ]_T$	1.79 (0.29)	1.80 (0.26)
	$[aɪ]_D$	1.47 (0.22)	1.50 (018)
Offglide (F2/F1)	$[aɪ]_T$	8.78 (0.76)	9.15 (1.47)
	$[aɪ]_D$	7.17 (1.31)	8.02 (1.75)

block 1, group A contributed the voiceless congruent stimuli, group B contributed the voiceless incongruent stimuli, group C contributed the voiced congruent stimuli, and group D contributed the voiced incongruent stimuli. This changed in the subsequent three blocks. In block 2, for example, group A contributed the voiced incongruent stimuli, group B contributed the voiceless congruent stimuli, group C contributed the voiceless incongruent stimuli, and group D contributed the voiced congruent stimuli. And so on. The purpose of this manipulation was to ensure that every participant heard all four versions (voiced versus voiceless, congruent versus incongruent) of the 14 pairs, but in different blocked orders.

The order of presentation of each version of each pair (voiced or voiceless consonant in congruent or incongruent condition) was varied across participants and lists in a Latin-Square design. Within each block, presentation of tokens was quasi-randomized such that participants never heard two members of the same word pair in the transition between blocks. Each list repeated the four blocks four times, such that participants heard and responded to four repetitions of each test token ($=224$).

Each list of test tokens was preceded by an exposure phase, during which participants heard one token of each target word ($=28$). The exposure words were the unmodified base tokens; that is, they were natural productions recorded from the same speaker. The purpose of the exposure phase was to create the expectation that, within the domain of the experiment, both members of a voicing pair were equally likely to occur, thereby mitigating the influence of any word-based bias for one response over the other. Data from the exposure tokens were not analyzed.

3.1.2. Procedure

Participants were run individually in a quiet laboratory room. They were seated in front of a computer screen and they wore high-quality headphones. At the beginning of the experiment, written directions instructed them how to complete the trials, i.e. by clicking on the word that corresponded to the word they had heard. On each trial, a 5×5 grid appeared on the screen with a central fixation point, and two printed words to the left and right of the point, as in Fig. 6. After 1 s, the spoken stimulus played. The grid remained on the screen until the participant clicked on one of the words. After a 1 s delay, the next trial began. One of the printed words corresponded to the target, and the other corresponded to its competitor. The target word was always defined in terms of the actual voicing qualities of the word's coda (regardless of whether the diphthong itself was congruent or incongruent), while the competitor had the opposite voicing quality. For example, for the spoken stimuli [baɪt]$_{\text{T-Short}}$ or [baɪt]$_{\text{T-Long}}$, the target was *bite* and the competitor was *bide.* For [baɪd]$_{\text{D-Long}}$ or [baɪd]$_{\text{D-Short}}$, the target was *bide* and the competitor was *bite.*

Beginning with onset of the spoken stimulus, participants' eye movements were recorded during the task using an EyeLink 1000 remote eye-tracker with a sampling rate of 500 Hz. The data were first parsed into fixations (i.e., looks at an object) and saccades (i.e., movements of the eyes from one object to another) using the eye-tracker's default algorithm (i.e., the onset of a saccade is marked when velocity is greater than 30°/s or when acceleration is greater than 8000°/s^2). Fixations were then automatically categorized as falling within one of three regions on the screen (i.e., the square where the target word was displayed, the square where its competitor was displayed, and the remaining area of the grid without further distinction). Note that because the eye-tracker corrects for head movements, our analysis does not differentiate between saccades which were accompanied by head movements and those which were not. On a given trial, recording of eye movements ended when the participant clicked on a printed word.

3.1.3. Participants

Twenty-five individuals from the University of Pennsylvania subject pool, none of whom had participated in Experiment 1, participated in the experiment for course credit. Data from three participants were excluded due to technical problems or difficulty in attaining acceptable quality with the eye-tracker calibration, leaving us with analyzable data from 22 participants.

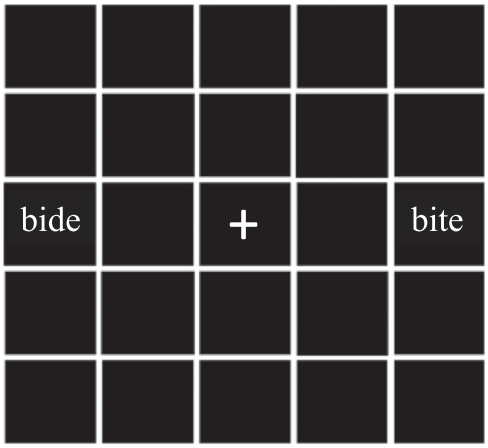


Fig. 6. Visual display for Experiment 2, showing fixation point, target word corresponding to spoken stimulus on a given trial (e.g., printed *bite* corresponds to [baɪt]$_{\text{T-Short}}$), and corresponding competitor word (e.g., printed *bide*).

3.1.4. Predictions

We predicted reduced target responses in the incongruent conditions ([baɪt]$_{\text{T-Long}}$, [baɪd]$_{\text{D-Short}}$), compared to the congruent conditions ([baɪt]$_{\text{T-Short}}$, [baɪd]$_{\text{D-Long}}$), which could manifest itself as decreased accuracy in the forced-choice identification task and/or delayed fixation curves in the eye-tracking data.

3.2. Results

3.2.1. Forced-choice responses

For the forced-choice task, we analyzed participants' error rates. A correct response was one in which coda voicing in the spoken stimulus corresponds to coda voicing on the visual target. For example, if the participant heard [baɪt]$_{\text{T-Short}}$ or [baɪt]$_{\text{T-Long}}$, where the coda was voiceless, the correct response was "bite". An error, then, is a response in which coda voicing in the spoken stimulus did not correspond to coda voicing on the visual target, in this case "bide".

Results show that errors were quite rare, averaging 1% of all responses, and equally distributed across conditions. These trials were excluded from further analyses. A multi-level logistic regression analysis, with Participant and Word Pair as random factors, revealed no significant main effects or interactions. (Appendix C shows the full analysis).

3.2.2. Eye fixations

Data from a given trial were excluded if (1) the participant's last (or only) mouse response was not to the target region (0.6% of the trials) or (2) the participant did not initiate a fixation to the target orthographic word during the window of analysis. This window started at the onset of the spoken diphthong (defined in the same manner as in the segmentation procedure for Experiment 1, see Section 2.1.4) and extended until the participant indicated their choice by clicking on one of the orthographic words. Instances in which no fixation to the target orthographic word were recorded prior to, or concomitant with, a click on the target may be due to reduced quality in the eye-tracker's calibration; alternatively, they may reflect the participant's choice not to move their gaze toward the object of their attention. In either case, the relationship between speech interpretation and direction of gaze is compromised in these trials, and it is thus deemed appropriate to exclude them trials from further analyses. This procedure excluded 4.6% of the trials, distributed equally across Voicing and Congruency conditions.

In order to capture changes in gaze behavior over time as a function of the stimulus that participants heard, we computed, for each participant and each condition, the proportion of trials for which a given word type or location (target, competitor, or the grid) was fixated during each successive 10-ms time window, from 0 to 1000 ms after the spoken diphthong onset. When, for a given trial, no orthographic word or location was fixated during a given 10 ms-time bin (because of an on-going blink or saccade), the trial was excluded from the total number of the trials over which the fixation proportion was computed for that time bin. In the event that the last fixation of the trial (always to the target orthographic word) ended before 1000 ms, its duration was extended. This procedure ensures that the trials in which a participant evidently identified the target word continue to contribute to the proportion of target fixations computed on the time bins remaining in the analysis window. This decision is motivated by our interpretation of fixation proportions in the context of this spoken-word-to-orthographic-word matching task: that the proportion of fixations to an orthographic word at a particular moment in time reflects the degree to which people consider that word to be the target. Clicking on one of the orthographic words (while or after fixating on it and without revising this choice) is evidence that the listener has identified the referent of the spoken word and the trial should be included in the proportion calculation, regardless of how long the final fixation lasted. Thus, the data are altered in a way that is entirely consistent with our interpretation of gaze behavior in the present task. An alternative to this procedure, one in which a trial whose last fixation to target has terminated before the end of the analysis window is excluded, is problematic in our view. Indeed, because the rate with which trials terminate over time is likely to differ across conditions, the procedure of removing terminated trials obscures our ability to observe a difference between the conditions because the trials that differentiate the conditions should be selectively excluded from the analysis.

The proportion of eye fixations to targets over time, averaged across participants for each condition, is depicted in Fig. 7. The four plotted curves begin to rise and diverge from one another around 200 ms, which corresponds to the typical delay in programming an eye movement (see Fischer (1987)). One curve is distinct from the others: namely, the curve from the voiceless congruent condition, which rises earlier than the other three curves and suggests that participants listening to [aɪ]$_{\text{T-Short}}$ stimuli rapidly directed their gaze to the target word. The curve from the voiceless incongruent condition rises more slowly, suggesting that participants listening to [aɪ]$_{\text{T-Long}}$ stimuli did not direct their gaze to the target word as rapidly. The separation between these two curves provides evidence that listeners' perception of words with voiceless codas relies, in part, on a short duration ratio – above and beyond the nucleus and offglide qualities that accompany such a ratio. Notice also that the voiceless congruent and incongruent curves exhibit similar slopes until about 300 ms: this is expected, if listeners' interpretations are initially influenced by nucleus quality, which was centralized in both cases. Around 300 ms, however, the voiceless congruent curve continues its steep rise while the voiceless incongruent curve exhibits a more shallow rise: this suggests that the delayed transition from nucleus to offglide in the incongruent condition mismatched listeners' original, voiceless interpretation of the stimulus. In other words, the extended duration of the nucleus in an [aɪ]$_{\text{T-Long}}$ stimulus interfered with its perception as a voiceless-coda word.

By contrast, the curves in the voiced condition do not differ according to the congruency or incongruency of the duration ratio, and both exhibit delayed rises when compared with the voiceless congruent condition. We propose a possible account for these findings in Section 3.3.

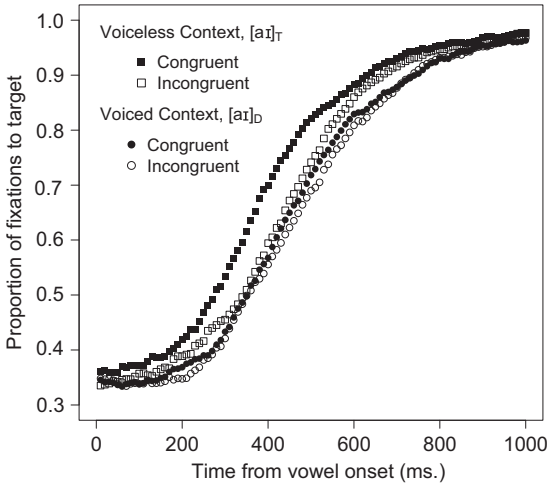


Fig. 7. Proportion of eye fixations to target words over time, for the four conditions of Experiment 2. Diphthongs from voiceless context = $[aɪ]_T$, from voiceless context = $[aɪ]_D$. Congruent stimuli are those for which the ratio [nucleus duration/offglide duration] is appropriate to the voicing context; i.e. short ratios for voiceless contexts and long ratios for voiced contexts. Incongruent stimuli are those for which the ratio [nucleus duration/total duration] is inappropriate to the voicing context; i.e. long ratios for voiceless contexts and short ratios for voiced contexts. Plotted data includes only those trials for which participants made correct responses. On *x*-axis, time 0 is aligned with the beginning of the diphthong.

In order to establish the statistical significance of the differences among the curves, the changes in target fixation proportions over time were analyzed using a growth-curve analysis technique (Mirman, Dixon, & Magnuson, 2008). Briefly, the goal of this kind of analysis is to characterize the curve (or function) formed by the series of fixation proportions over time by a small set of parameters. This approach is akin to capturing the values of a variable (*y*) over values of another variable (*x*) that fall on a straight line by two parameters, the intercept and the slope of the line. Here, because the relationship between fixation proportion and time is much more complex than a straight line, we modeled it using power polynomials, as implemented by the poly() function in the stats package of R. Power polynomials can capture the curvilinear relationship between fixation proportion and time by raising the variable 'time' to a particular power. When a curve is characterized by two inflection points, as is the case in the curves we model here, the function can be well approximated by raising the variable 'time' to the power of 3.

Unfortunately, in power polynomials, the different terms are not independent from one other: adding a higher-order term, such as the cubic terms, affects the estimates associated with the lower-order terms. This is an issue because the parameters and their values become dependent on the experimenter's choice of the curve complexity. Because we have no independent reasons to assume that the curve we want to model must be linear, quadratic, or cubic, the experimenter's choice has a disproportionate impact on the outcome. To circumvent the issue, and following Mirman et al.'s (2008) recommendations, we use orthogonal power polynomials. With orthogonal polynomials, each term is established independently from the others. For instance, the estimate of the linear term (i.e., the term that models the changes of fixation proportion over time as a straight line) remains the same whether or not a quadratic term (which captures the presence of an inflection point in the curve) is added. Finally, the growth-curve analysis also allows us to specify and model the clustering of the data in terms of fixed and random effects (see Mirman et al. (2008), for a helpful discussion on the hierarchical component of the method).

The first step is to establish a base model that captures how fixation proportions varied over time when averaged across all conditions. Such a model includes aspects of the curve that may differ across participants for reasons that do not concern us here. Here, we computed a base model with four orthogonal time terms (i.e., intercept, linear, quadratic, and cubic), incorporating random effect terms on the intercept and linear terms for each participant. (Attempts to model random effects on the quadratic or cubic terms resulted in models that failed to converge).

In a second step, the base model is augmented by terms (or predictors) that capture the effect of our predictor variables. The rationale is to test whether an augmented model approximates the data significantly better than the base model, taking into account the additional degrees of freedom that augmented models benefit from. This, in essence, tells us whether a given factor (or interaction between factors) significantly affects the fixation proportion curves. For Experiment 2, we successively added terms associated with the effect of coda voicing and of ratio congruency, as well as their interaction, and we compared goodness of model fit with the immediately simpler model. The difference in goodness of fit was measured as -2 times the change in log likelihood, which follows a χ^2 distribution with degrees of freedom equal to the number of parameters that were added to the simpler model. This procedure enables us to assess the significance of each term's contribution to fitting the observed data.

Table 7 presents results from model comparisons. As the values for $Pr(>\chi^2)$ show, models that incorporated the contribution of the ratio congruency on the function intercept accounted for the data significantly better than the base model that did not incorporate this factor, even though they introduced one additional degree of freedom (as indicated by the (1) in $\chi^2(1)$). Furthermore, the addition of this factor on each of the time terms improved the model's fit significantly; this indicates that the factor influenced the overall shape of the curves, not just the size of the area under the curve (as indicated by the intercept) or the slope of a line that best fits the curve (as indicated by the linear term). Incorporating the contribution of the factor voicing on each of the time terms also improved the models'

Table 7
Growth curve analysis for eye fixations, Experiment 2. See text for explanation.

Model	Log likelihood	$\chi^2(1)$	Pr(>χ^2)
Base	*−50519*		
Main effects			
Ratio congruency (congruent versus incongruent)			
Intercept	−50412	213.6	<0.0001
Linear	−50399	27.4	<0.0001
Quadratic	−50355	87.5	<0.0001
Cubic	−50342	26.3	<0.0001
Voicing (voiceless vs. voiced)			
Intercept	−50129	781.2	<0.0001
Linear	−50120	17.4	<0.0001
Quadratic	−50008	224.6	<0.0001
Cubic	−49997	20.8	<0.0001
Model with main effects	*−49797*		
Two-way interaction			
Matching:Voicing			
Intercept	−49747	100.3	<0.0001
Linear	−49738	16.4	<0.0001
Quadratic	−49727	22.9	<0.0001
Cubic	−49719	16.3	<0.0001

fit. Importantly, models that included terms related to the interaction between voicing and congruency fitted the data significantly better than a model that included only their main effects.

In the full model, areas under the curve are overall smaller in the voiced condition, compared to the voiceless condition which served as the baseline ($\beta = -57.42$, std. error = 2.05). Areas under the curve are also overall smaller in the incongruent condition, compared to the congruent condition which served as the baseline ($\beta = -37.28$, std. error = 2.05). Unlike the parameter estimates for the main effects, the estimate for the interaction between voicing and congruency has a positive sign ($\beta = 28.98$, std. error = 2.90), indicating that the effect of incongruency is smaller in the voiced condition, compared to voiceless.

These analyses confirm much of what is apparent in Fig. 7. The main effects show that ratio congruency significantly affected participants' fixations to target orthographic words, as did coda voicing context. Furthermore, the interaction shows that congruency had a significantly larger effect when the diphthong came from a voiceless coda context, compared to when it came from a voiced one.

3.3. Summary

Experiment 2 provides some support for the perceptual predictions of our gestural timing proposal. When listeners heard congruent stimuli from voiceless contexts ([aɪ]$_{\text{T-Short}}$), they looked at the corresponding visual target earlier and more often than when they heard the incongruent stimuli ([aɪ]$_{\text{T-Long}}$). Crucially, they did so despite the fact that [aɪ]$_{\text{T-Short}}$ and [aɪ]$_{\text{T-Long}}$ contained equivalent overall durations, nucleus qualities, and offglide qualities. This is evidence that listeners are sensitive to the duration of the nucleus (either in absolute terms, or relative to the diphthong's total duration). For example, in a voiceless coda environment where a representative overall diphthong duration is 224 ms (see Fig. 5), a nucleus that extends for a contextually inappropriate value of 87 ms (rather than an appropriate value of 38 ms) is perceived to be "too long" and interferes with identification of the target word as ending with a voiceless consonant. This finding, we argue, demonstrates that listeners are sensitive to the timing of gestures involved in the production of the diphthong.

The lack of congruency effect in the voiced context may appear to question this conclusion. However, close inspection of the stimuli and results suggests a possible explanation. As our results show, listeners make many fixations before they have heard the entire spoken word, or even the entire vowel. Thus, their decisions about which word they think they hear, particularly early on, are likely to be driven by absolute durations of the nucleus, rather than by its duration relative to the entire vowel. As shown by the representative examples in Fig. 5, the absolute nucleus durations for incongruent stimuli were quite close across conditions: 87 ms for the voiceless incongruent condition, and 80 ms for the voiced incongruent condition. As our results indicate, a nucleus duration of 87 ms interferes with identification of voiceless-coda targets in a forced-choice situation. Logically, once listeners have decided that this value has become "too long" for a voiceless-coda word, they may shift their interpretation to the voiced-coda competitors that were present on the visual display, for which long nucleus durations are appropriate. However, our results also indicate that a nucleus duration of 80 ms does *not* interfere with identification of voiced-coda targets: while this value is technically "too short" for voiced-coda words, it is also "too long" to activate the voiceless-coda competitors. Thus, in a forced-choice situation, a nucleus duration of 80 ms – even if it is an unnatural value for voiced-coda targets in laboratory speech – does not drive fixations to the voiceless-coda competitors. Presumably, a much shorter value, such as the 39 ms nucleus duration used in the voiceless congruent condition, would be required to produce such interference in the voiced incongruent condition. Further work with a greater range of duration ratios could confirm this scenario.

An aspect of the data that is particularly intriguing concerns the difference in fixation curves to the voiced versus voiceless targets in the congruent conditions. When listeners heard voiceless congruent stimuli ([aɪ]$_{\text{T-Short}}$), they looked at the target earlier and more often than when they heard voiced congruent stimuli ([aɪ]$_{\text{D-Long}}$). In other words, [aɪ]$_{\text{T}}$ seems to have higher predictive value for T than [aɪ]$_{\text{D}}$ has for D. This finding was unexpected, and not specifically predicted by our hypothesis. Why might voiced-coda words be identified more slowly than voiceless-coda words?

One possible explanation comes from the phenomenon of raising, in which /aɪ/$_{\text{T}}$ alternates to [ɜi] or [ʌi]. Canadian raising occurs in many – but crucially, not all – northern dialects of American English (see Moreton and Thomas (2007)). The participants in Experiment 2, who were all undergraduates at a northern U.S. university which enrolls students from many dialect regions, most likely have regular exposure to both raising and non-raising dialects. In other words, they have regular exposure to both [aɪ] and [ɜi] occurring before a voiceless consonant. Meanwhile, because Canadian raising is triggered only by voiceless environments, the participants have exposure only to [aɪ] before voiced consonants. This pattern creates an asymmetry in the predictive value of [aɪ] versus [ɜi] tokens. When a listener hears a token that sounds more like [aɪ], with a peripheralized nucleus and centralized offglide, the following consonant could be either voiced or (if listening to a non-raising speaker) voiceless. However, when a listener hears a token that sounds more like [ɜi], with a centralized nucleus and peripheralized offglide, the following consonant *must* be voiceless (and the speaker must have a raising dialect) because voiced consonants do not trigger this realization. Some nucleus raising may occur even in non-Canadian-raising dialects (Thomas, 2000), however, so this explanation would need to be confirmed by further research. Also, the scope of this explanation does not extend beyond the diphthong /aɪ/; if it is correct, we would not expect to see similar asymmetries in other diphthongs or monophthongs.

Another explanation for the voicing effect is wider in scope, and makes broader predictions: it could be the case that temporal proximity of adjacent gestures produces signatures in the acoustic record that provide listeners with richer cues to upcoming segments than the signatures produced by temporal separation. This explanation would fit with previous findings in the literature on perception of co-articulation, which show that co-articulated segments act as strong predictors of upcoming segments, while plain segments (with no co-articulation) act as relatively weak predictors (Beddor et al., 2013; Flagg, Oram Cardy, & Roberts, 2006; Salverda et al., 2014). In studies examining vowel nasalization, for example, Beddor et al. (2013) and Flagg et al. (2006) demonstrate that while co-articulated $\tilde{\text{V}}$ is a strong perceptual predictor of upcoming nasal N, plain V is a comparatively weak predictor of an upcoming oral C. If this finding generalizes to more cases of gestural timing before codas, [aɪ]$_{\text{Voiceless}}$ should have higher predictive value for voiceless codas than [aɪ]$_{\text{Voiced}}$ has for voiced ones, as our current results seem to suggest. Future research could test this possibility further.

To summarize, our results offer support for our proposed gestural basis for vowel duration and spectral differences. In voiceless coda contexts, incongruent diphthongs produced later and fewer looks to corresponding visual words, despite the fact that all other diphthong cues for voicing – overall duration, nucleus quality, and offglide quality – remained congruent. This effect suggests that listeners were sensitive to the conflict between duration ratio and the other cues; put differently, it suggests listeners were sensitive to the gestural origins of the durational and spectral cues they were listening to.

4. Discussion

4.1. Summary of findings

Overall, our results for the American English diphthong /aɪ/ provide support for the idea that coda voicing triggers changes in the relative timing of articulatory gestures. Our first prediction concerning speech production was clearly confirmed. In Experiment 1, the acoustic measure that we used to index temporal overlap and displacement between the gesture for nucleus [a] and the gesture for offglide [ɪ], namely the ratio between nucleus duration to offglide duration, showed significantly different patterns in voiceless versus voiced contexts, and these differences remained robust across conditioning contexts and individual speakers. Our second prediction concerning speech perception was also partially confirmed. In Experiment 2, diphthong stimuli in which acoustic cues were congruent with their hypothesized gestural origins facilitated listeners' responses to targets compared to stimuli in which cues were incongruent, although this finding held only for diphthongs in voiceless coda contexts.

4.2. Limitations

One limitation of the current study concerns its focus on acoustic data, rather than kinematic data. In interpreting the production findings of Experiment 1, then, we relied on an acoustic index of temporal alignment (the ratio of [nucleus duration/offglide duration] in [aɪ]), rather than a more direct indicator. In setting up the perceptual stimuli for Experiment 2, we relied on this same acoustic indicator. Our methodological decision is consistent with a coarse-grained view of gestures as theoretical primitives: that is, we evaluated the extent to which the acoustic record and perceptual patterns were consistent with our proposed gestural scores. For Experiment 1, this decision also allowed us to collect data for a larger number of participants than are typically collected in studies using, e.g., electromagnetic articulometry. We were therefore able to investigate an important prediction of our hypothesis, namely that gestural timing patterns before voiceless and voiced codas should be consistent, not variable, across multiple speakers. Nevertheless, in order to more fully test a theory about gestures, we would need production data to provide us with a finer-grained

view of those gestures, namely kinematic data such as that provided by articulometry or ultrasound; and we would need perceptual stimuli generated by models of constriction location and degree within the vocal tract.

A second limitation of our approach is the reliance on speech produced in a laboratory setting using sentences that are unlikely to occur in everyday situations. Thus, the speech that our participants produced and perceived may differ from what they would encounter in natural settings. An advantage of our approach, however, is that we were able to control the variables of interest.

Another limitation of this study concerns its focus on a single English diphthong, /aɪ/. In order to generalize our results, we must expand our predictions to other diphthongs. Previous researchers have already laid some of this groundwork. Gay (1968), Lehiste and Peterson (1961), and Moreton (2004) all report production patterns for other English diphthongs that mimic those found in /aɪ/. Moreton (2004) also reports perception patterns for /eɪ/ that mimic those found in /aɪ/, and concludes that "there is nothing special about /aɪ/ in either production or perception" (2004: 24). Given these findings, it seems likely that our proposed account could apply straightforwardly to other diphthongs.

4.3. Expanding to monophthongs

Future work should also expand the gestural timing hypothesis to monophthongs. This would require formulating gestural scores for words like *pot* [pɑt] and *pod* [pɑd], in which a single tongue-body gesture precedes a voiceless or voiced coda, and testing the predictions made by these scores. Fig. 8 sketches one possibility.

As with the gestural scores that we proposed for diphthongs in Fig. 1, the scores for monophthongs in Fig. 8 propose that duration changes in coda voicing environments crucially occur along with changes in relative gestural timing. In voiceless contexts as in *pot* [pɑt] (top panel), the tongue-body gesture for /ɑ/ has a relatively short inherent duration, as expected, but also shifts leftward in time. As a consequence, it would be relatively separated from the tongue-tip gesture for /t/, such that it has ample time to reach its articulatory target without the potential for overlap. In voiced contexts as in *pod* [pɑd] (bottom panel), the tongue-body gesture for /ɑ/ has a relatively long inherent duration, but also shifts rightward in time. As a consequence, it would be relatively close to the tongue-tip gesture for /d/, which overlaps with it.

The dashed lines in Fig. 8 show plausible acoustic segmentations that would be produced by these gestural scores: *y* indicates the beginning of /ɑ/ and *z* shows the end of /ɑ/. We have vertically aligned the top and bottom panels according to *z,* in order to visually orient the reader toward the differences between them. Because co-articulatory overlap can occur between the gesture for /ɑ/ and the gesture for the coda, in a manner that we propose is modulated by the voicing of the coda, *z* occurs in different locations in each panel: in the top panel, with no overlap, it occurs equidistant from the end of /ɑ/ and the beginning of /t/. In the bottom panel, with overlap, it occurs at the beginning of /d/ (We discuss the dotted line and *y′* in Section 4.4).

Compared to the *pod* configuration in the bottom panel of Fig. 8, then, the *pot* configuration could conceivably manifest itself in the acoustic record as (a) relatively short overall duration of the monophthong, and (b) relatively peripheralized formants for the vowel [ɑ]. In the production domain, the predictions associated with the gestural scores in Fig. 8 have already been demonstrated in previous acoustic measurements for vowel duration, which are shorter in voiceless environments and longer in voiced environments (House & Fairbanks, 1953; Peterson & Lehiste, 1960) as well as for low vowel spectral qualities, which are peripheralized in voiceless environments and centralized in voiced ones (references cited in Moreton, 2004). Future work would need to fill the gap in the literature on high vowel spectral qualities. In the perception domain, the duration ratio of internal vowel components is not an operative concept for monophthongs in the same way that it is for diphthongs, so it would be difficult to create stimuli that manipulated such a variable. However, as we discuss below, the gestural scores in Fig. 8 also make predictions about onset-to-vowel and vowel-to-coda durations; manipulating these ratios could test listeners' sensitivity to the gestural origins of duration changes in monophthongs.

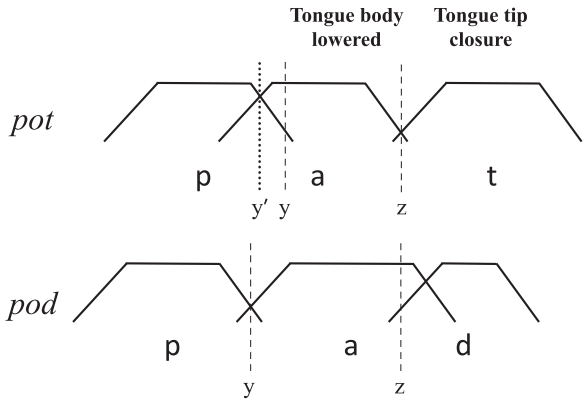


Fig. 8. Possible gestural scores for words with monophthongal vowels before voiceless codas, as in *pot* [pɑt] (upper panel) and before voiced codas, as in *pod* [pɑd] (lower panel). For clarity, the glottal gesture for the coda is omitted. Dashed lines indicate acoustic segmentation. Dotted line indicates an alternative acoustic segmentation (see text).

4.4. Additional predictions for onsets and codas

In addition to the predictions they make for vowel duration and spectral properties, the gestural scores in Fig. 8, as well as those proposed by Beddor (2009) for CVNC words, make some predictions for the properties of onset and coda consonants. To see this, we will refer to the difference in location between *y* and *y′* in the top panel. *y* indicates the beginning of [ɑ] in the acoustic record, specifically when it is preceded by an onset gesture whose articulator is independent of the tongue body used for the vowel (e.g. the lips for the /p/ in *pot,* on the assumption that the labial closure hides some portions of the following tongue-body constriction, and the aspirated release from closure also overlaps with it somewhat). Comparing the location of *y* across the top and bottom panels of Fig. 8, then, we see that the acoustic duration of the onset should be equivalent across voicing conditions.

y′, on the other hand, indicates the beginning of [ɑ] in the acoustic record, specifically when it is preceded by an onset gesture produced with a non-independent articulator (e.g., the tongue body for /k/ in *cot* and *cod,* on the assumption that the closure at the velum for /k/ must transition directly to a lowered position for /ɑ/, with no overlap possible). Comparing *y′* in the top panel versus *y* in the bottom panel, we see that the acoustic duration of the onset should be shorter in voiceless coda environments, and longer in voiced ones. In other words, the prediction is that onset consonants produced with the tongue body should be acoustically shorter in words like *cot,* and longer in words like *cod.* In fact, at least one previous study offers some support for this prediction. In a speech corpus study, Hawkins and Nguyen (2004) found that onset /l/ exhibits shorter acoustic durations when it occurs in words with voiceless codas, than when it occurs in words with voiced codas. Future work would need to demonstrate a similar pattern for other onsets.

Although our proposed gestural scores conceivably would make analogous predictions for acoustic duration of coda consonants, these predictions are more difficult to evaluate given the known differences in duration between voiceless versus voiced coda consonants, which plausibly arise from a different source, such as the aerodynamic voicing constraint (Ohala, 1997). However, Beddor (2009) points to a related prediction: *within* a given voicing category, shorter vowel durations – either for the offglide portion of a diphthong, or for an entire monophthong – should co-occur with longer coda durations. This is because, by hypothesis, earlier occurrence of the coda gesture is what clips the preceding vowel gesture more, while later occurrence clips it less. Analogously in the case of CVNC words, shorter nasal consonants should co-occur with longer coda consonants, and Beddor (2009) provides data demonstrating that, within the voiceless coda category, this prediction holds. Unfortunately, it was not possible to test these predictions in our own data because our participants produced highly variable realizations of coda /t/ and /d/, particularly under fast speech conditions, and so we must leave this prediction to future work.

We could also further test our proposal for gestural timing by examining place-of-articulation effects in both onsets and codas. As Figs. 1 and 8 show, we have proposed temporal proximity between the gestures for onsets and vowels in voiceless environments, and temporal proximity between the gestures for vowels and codas in voiced environments. Importantly, the same notion of "proximity" between two adjacent gestures for a vowel and a consonant predicts different consequences, depending upon the consonant's place of articulation. This is because, while the gesture for a vowel always involves the tongue body, the gestures for an adjacent consonant may involve the same articulator (that is, tongue body for velars such as /k/ and /g/), a somewhat independent articulator (tongue tip for alveolars such as /t/ and /d/), or a completely independent articulator (lips for labials such as /p/ and /b/). Taking vowel-coda sequences as an example, then, a vowel-velar sequence would involve a transition from a tongue-body constriction in one location, to a tongue-body constriction in another location: because the same articulator is used for each constriction, there is no potential for gestural overlap, even when the gestures are close to one another in time. By contrast, a vowel-labial sequence would involve a transition from a tongue-body constriction to a labial constriction: because different, independent articulators are used, there is indeed potential for gestural overlap, as long as the gestures are close in time. The overall picture that emerges is one in which gestures for voiced velar codas clip the preceding vowel gesture to a relatively small degree, while gestures for voiced labial codas clip it relatively large degree, with gestures for voiced alveolar codas somewhere in between. A similar logic could apply to onset-vowel sequences. This sets up a rich set of predictions for future research.

Results from a previous perceptual study already offer some tentative support for these ideas. Hillenbrand et al. (1984) manipulated the nonsense monosyllables [pɛb, pɛd, pɛg, pag, pig, pug] by cutting acoustic information from the end of each syllable in successive increments of 10 ms, and asked listeners to make coda voicing judgments. Their basic finding was that voiceless coda responses were not elicited in large numbers until the closure interval and, in most cases, a portion of the vowel-coda transition had been removed. As the authors state, "[t]his finding is consistent with the idea that information in the VC transitions is important to the perception of final voicing contrast" (1984: 21). Unexpectedly, however, their results showed significant place of articulation effects. Stimuli with labial codas, /pVb/, were judged to be voiceless *earlier* (that is, with *less* acoustic information cut from end of word) than stimuli with non-labial codas, /pVd/ and /pVg/. The authors do not offer an explanation for this effect, but gestural timing can potentially provide one. If the gestures for vowels and adjacent voiced codas are always close to one another, as we propose, then in a Vb sequence, the labial gesture for [b] can overlap significantly with the preceding tongue-body gesture for V, potentially clipping it to a large degree. In a Vd or Vg sequence, by contrast, the tongue-tip or tongue-body gestures can overlap only somewhat or not at all with the preceding tongue-body gesture for V, clipping it relatively little. These different clipping effects should manifest themselves in the acoustic record. On this analysis, it makes sense to think listeners would judge /pVb/ stimuli to be voiceless earlier because, in a sense, an overlapping Vb sequence has naturally begun the very manipulation that Hillenbrand et al. (1984) artificially performed, namely clipping temporal information from the end of the vowel.

4.5. Comparison to other theories

4.5.1. Theory of articulatory undershoot

Lindblom's (1963) theory of articulatory undershoot predicts that shorter vowel durations should correlate with spectral centralization. On its own, therefore, this theory does not adequately account for coda voicing environments in American English, where we see the inverse correlation. On the other hand, the gestural timing analysis we have proposed still relies upon a basic tenet of undershoot, namely that gestures require a certain amount of time to fully achieve their targets. For example, we rely upon the idea of increased duration for the offglide /ɪ/ gesture, and hence full achievement of the target articulation, in order to account for the fact that the acoustic qualities of offglide [ɪ] are peripheralized in voiceless environments. The difference is that, in our proposal, increased duration is not an isolated statement about the /ɪ/ gesture, but a relativized statement that crucially relies upon the timing of the preceding gesture for /a/ and the subsequent gesture for the coda. In this sense, our proposal does not replace the theory of articulatory undershoot, but elaborates upon it by enriching the inventory of gestural timing relationships that can occur. This is consistent with the spirit of previous work on phrasal lengthening, which has shown that speakers can use multiple different timing relationships to accomplish changes in duration (Byrd, 2000; Edwards et al., 1991; Krivokapić & Byrd, 2012).

4.5.2. Theory of lengthened nuclei

In a production study of American English diphthongs, Gay (1968) concluded that "[t]he increased duration of /ɔɪ, aɪ/ preceding a voiced consonant is accomplished primarily by a lengthening of the steady-state onset." As a statement of empirical findings, this conclusion is fully compatible with the results of the current study, as well as previously-reported results (Jacewicz et al., 2003; Lehiste & Peterson, 1961; Moreton, 2004; Thomas, 2000). As a theoretical proposal, however, this conclusion would not account for the spectral changes that occur in offglide [ɪ] without further stipulation. Nor would it extend straightforwardly to monophthongal vowels, which do not have straightforward analogs to nuclei ("steady-state onsets") and offglides. With that said, the notion of lengthened nuclei does capture something correct about the CVNC words that Beddor (2009) investigates, provided we consider the non-nasalized portions of the monophthongal vowel in a word like *bend* [bɛ̃nd] to be analogous to the nucleus of the diphthong in a word like *bide* [baɪd]. In both cases, the "nucleus" (either [a] or the non-nasalized portion of [ɛ]) lengthens before a voiced coda. Like Gay's statement of lengthened nuclei, our proposal for gestural timing also captures the CVNC pattern; unlike his statement, however, our proposal is rich enough to offer a unified account of the durational and spectral changes occurring in diphthongs, and potentially monophthongs as well.

4.5.3. Theory of hyper-articulation

Moreton (2004) suggests that spectral changes in vowels are due not to timing relationships, but to the specific voicing properties of the coda. Under his account, voiceless obstruents are hyper-articulated – that is, pronounced with more force – than voiced obstruents, and this hyper-articulation undergoes regressive spread to the preceding vowel, which therefore become peripheralized compared to those in voiced obstruent environments. In support of this concept, Moreton reports production data demonstrating peripheralization of both F1 and F2 for the offglides of English diphthongs [aɪ, ɔɪ, eɪ, aʊ] before voiceless codas. He also reports perception data demonstrating that increasingly peripheralized F1 and F2 in the offglides of [aɪ] and [eɪ] facilitate identification of voiceless codas, as do shorter nucleus durations. Note that both sets of results are entirely compatible with the gestural timing account that we propose, and with our own production and perception data.

Both our gestural timing proposal and Moreton's hyper-articulation proposal can extend to monophthongal cases in a reasonably straightforward manner: for gestural timing, this would mean adopting something along the lines of the gestural scores in Fig. 8, while for hyper-articulation, this would mean stating that hyper-articulation undergoes the same type of regressive spread for preceding monophthongs as it does for diphthongs. Furthermore, both proposals can account for the fact that, in Moreton's data as well as in previous production studies, peripheralization does not affect all parts of the vowel equally, but occurs primarily at the vowel offset near voiceless coda. According to gestural timing, this pattern would occur because the gesture for the voiceless coda is close in time to the right edge of the gesture for the preceding vowel; according to hyper-articulation, it occurs because forcefulness, somewhat like traditional phonological features, spreads only locally.

Nevertheless, the differences between these two proposals are substantial. Moreton's (2004) hyper-articulation account does offer one advantage, namely that it links spectral changes directly to the qualities of voicelessness per se, which gives some phonetic grounding the fact that diphthong hyper-articulation in voiceless environments recurs in languages of the world besides American English (Moreton, 2004: 3). As currently formulated, our gestural timing proposal cannot yet forge this link, and the co-occurrence of coda voicelessness with a particular gestural score, versus coda voicing with a different gestural score, remain unexplained. On the other hand, Moreton's proposal does not link spectral changes directly to duration changes. Nor does it straightforwardly extend to the CVNC cases discussed by Beddor (2009), the /l/ onset cases reported by Hawkins and Nguyen (2004), or the place-of-articulation perceptual effects reported by Hillenbrand et al. (1984). The gestural timing account can potentially capture all of these patterns, suggesting that it may be a more generalizable solution.

Acknowledgments

For helping us to conduct this research, we thank Adrian Benton and Shannon Fouse. For help with stimulus recordings, we thank Michael Key. For valuable feedback, we thank Ann Bunger, Sarah Johnstone Drucker, John Kingston, Tanya Kraljic, Tamara Nicol Medina, and audiences at the annual meetings of Psychonomics Society and of Architectures and Mechanisms of Language Processing. We are deeply grateful for the detailed comments provided by three anonymous reviewers and by editor Kenneth de Jong. This work was supported by a research grant from the National Institutes of Health (R01 HD 049742-1), and partially supported by an Integrative Graduate Education and Research Traineeship grant from the National Science Foundation (NSF-IGERT 0504487).

Appendix A

Results of mixed-effects linear models on four outcome variables in Experiment 1.

A.1 Total duration

	Estimate	Std. error	*t*-Value	*P*-value
(Intercept)	107.88	6.04	17.85	0.00
Voicing	6.46	3.47	1.86	0.06
PhrasalPos	38.94	3.47	11.21	0.00
Rate	55.55	3.47	15.99	0.00
Voicing:PhrasalPos	25.68	4.91	5.23	0.00
Voicing:Rate	43.28	4.91	8.81	0.00
PhrasalPos:Rate	−13.99	4.91	−2.85	0.00
Voicing:PhrasalPos:Rate	−4.08	6.95	−0.59	0.56

A.2 Duration ratio

log (Nucleus duration/Offglide duration)

	Estimate	Std. error	*t*-Value	*p*-value
(Intercept)	−1.67	0.14	−12.21	0.00
Voicing	0.33	0.13	2.56	0.01
PhrasalPos	0.05	0.13	0.39	0.70
Rate	−0.43	0.13	−3.33	0.00
Voicing:PhrasalPos	0.11	0.18	0.60	0.55
Voicing:Rate	0.53	0.18	2.88	0.00
PhrasalPos:Rate	−0.06	0.18	−0.33	0.74
Voicing:PhrasalPos:Rate	−0.19	0.26	−0.73	0.46

A.3 Nucleus

log (*F2/F1*)

	Estimate	Std. error	*t*-Value	*P*-value
(Intercept)	0.91	0.02	53.44	0.00
Voicing	−0.10	0.01	−8.51	0.00
PhrasalPos	−0.06	0.01	−5.04	0.00
Rate	−0.13	0.01	−10.65	0.00
Voicing:PhrasalPos	−0.06	0.02	−3.25	0.00
Voicing:Rate	−0.08	0.02	−4.56	0.00
PhrasalPos:Rate	0.03	0.02	1.94	0.05
Voicing:PhrasalPos:Rate	0.04	0.02	1.71	0.09

A.4 Offglide

log (F2/F1)

	Estimate	Std. error	*t*-Value	*p*-value
(Intercept)	1.44	0.02	62.65	0.00
Voicing	−0.07	0.01	−5.34	0.00
PhrasalPos	0.19	0.01	14.38	0.00
Rate	0.25	0.01	18.68	0.00
Voicing:PhrasalPos	−0.05	0.02	−2.64	0.01
Voicing:Rate	−0.07	0.02	−3.82	0.00
PhrasalPos:Rate	−0.10	0.02	−5.21	0.00
Voicing:PhrasalPos:Rate	0.03	0.03	0.95	0.34

Appendix B. Procedure for duration manipulation for Experiment 2 stimuli

To manipulate the ratio [nucleus duration/offglide duration] in the stimuli for Experiment 2, we used the Duration Manipulation commands in Praat, as follows. First, we converted each base sound file into a Praat "Manipulation" object. Then, we used a script to create a new Duration Tier for this object. The script added points on the Tier using the "Add Point" command. For example, to change a base nucleus from 65 ms to the target of 38 ms, we added a point at 0 ms (onset of the vowel) and another point at 65 ms; for each of these two points, we indicated the relative duration value of 38/65, which decreased the duration between the points by a factor of 0.58. Similarly, to change a base offglide from 159 ms to the target of 186 ms (in a vowel with total duration of 224), we added a point at 65.1 ms (to distinguish it from the other point at 65 ms) and another point at 224 ms. For each of these points, we indicated the relative duration value of 186/159, which increased the duration between these two points by a factor of 1.17. After running the script on a sound file, we manually selected "Replace Duration Tier" in Praat, opened the Manipulation object, and selected "Publish Resynthesis".

To manipulate duration of speech signals, Praat implements a method called Time-Domain Pitch-Synchronous Overlap-and-Add (TD-PSOLA™). As stated in the Praat help files and re-summarized here, creating a Sound from a Manipulation object involves four steps. First, a method called PitchTier: To PointProcess generates new points along the time domain. Second, Praat removes all points that lie within voiceless intervals from the new pulses, by using the period information in the original pulses (which is available in the Praat Manipulation object). Third, Praat copies the voiceless parts from the source to the target Sounds, while re-using some parts if the local duration is greater than 1. Fourth, for each target point, Praat looks up the closest source point. It copies a piece of the source Sound, centered around the source point, to the target Sound at a location determined by the target point, using a bell-shaped window whose left-hand half-length is the minimum of the left-hand periods adjacent to the source and target points (and analogously for the right-hand half-length). For more detailed information, the Praat help files refer the reader to Moulines and Charpentier (1990).

Appendix C. Statistical analysis of identification data for Experiment 2

Results of multi-level logistic regression analysis on Experiment 2 forced-choice identification results, with errors as the outcome variable, as implemented with the lmer() function in the lme4 package for R. Within each set of parentheses in the table, the first condition listed served as the baseline in the model. Participant and Word Pair were entered as random factors affecting the intercept.

	Estimate	Std. error	*z* Value	*Pr*(>*z*)
(Intercept)	−6.54	0.54	−12.00	<2e−16
Voicing (*Voiceless*)	−0.13	0.37	−0.34	0.74
Congruency (*Congruent*)	0.34	0.35	0.96	0.34
Voicing: Congruency	0.61	0.50	1.22	0.22

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Beddor, P. S. (2009). A coarticulatory path to sound change. *Language*, *85*(4), 785–821.

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., & Brasher, A. (2013). The time course of perception of coarticulation. *The Journal of the Acoustical Society of America*, *133*(4), 2350–2366.

Boersma, P. & Weenink, W. 2012. Praat: doing phonetics by computer (Version 5.3.29). ⟨⟨http://www.praat.org/⟩⟩.
Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica*, *49*(3–4), 155–180.
Byrd, D. (1992). Perception of assimilation in consonant clusters: a gestural model. *Phonetica*, *49*(1), 1–24.
Byrd, D. (2000). Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica*, *57*(1), 3–16.
Byrd, D., Kaun, A., Narayanan, S., & Saltzman, E. (2000). Phrasal signatures in articulation. *Papers in Laboratory Phonology V*, 70–87.
Cohn, Abigail C. (1990). Phonetic and phonological rules of nasalization. *UCLA working papers in phonetics*, vol. 76 (pp. 1–224).
Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367.
Denes, P. (1955). Effect of duration on the perception of voicing. *The Journal of the Acoustical Society of America*, *27*(4), 761–764.
Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *The Journal of the Acoustical Society of America*, *89*(1), 369–382.
Fischer, B. (1987). The preparation of visually guided saccades. In *Reviews of physiology, biochemistry and pharmacology*, Vol. 106 (pp. 1–35). Berlin: Springer.
Fischer, R. M., & Ohde, R. N. (1990). Spectral and duration properties of front vowels as cues to final stop-consonant voicing. *The Journal of the Acoustical Society of America*, *88*(3), 1250–1259.
Flagg, E. J., Oram Cardy, J. E., & Roberts, T. P. (2006). MEG detects neural consequences of anomalous nasalization in vowel–consonant pairs. *Neuroscience Letters*, *397*(3), 263–268.
Fourakis, M. (1991). Tempo, stress, and vowel reduction in American English. *The Journal of the Acoustical society of America*, *90*(4), 1816–1827.
Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, *68*(2), 161–177.
Gafos, A. I. (2002). A grammar of gestural coordination. *Natural Language & Linguistic Theory*, *20*(2), 269–337.
Gay, T. (1968). Effect of speaking rate on diphthong formant movements. *The Journal of the Acoustical Society of America*, *44*(6), 1570–1573.
Gay, T. (1978). Effect of speaking rate on vowel formant movements. *The Journal of the Acoustical Society of America*, *63*(1), 223–230.
Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: a phonology for public language use. *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, 159–207.
Gottfried, M., Miller, J. D., & Meyer, D. J. (1993). Three approaches to the classification of American English diphthongs. *Journal of Phonetics*, *21*(3), 205–229.
Hawkins, S., & Nguyen, N. (2004). Influence of syllable-coda voicing on the acoustic properties of syllable-onset /l/ in English. *Journal of Phonetics*, *32*(2), 199–231.
Hillenbrand, J., Ingrisano, D. R., Smith, B. L., & Flege, J. E. (1984). Perception of the voiced–voiceless contrast in syllable-final stops. *The Journal of the Acoustical Society of America*, *76*(1), 18–26.
House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, *25*(1), 105–113.
Iskarous, K. (2005). *Patterns of tongue movement. Journal of Phonetics*, *33*(4), 363–381.
Iskarous, K., & Kavitskaya, D. (2010). The interaction between contrast, prosody, and coarticulation in structuring phonetic variability. *Journal of Phonetics*, *38*(4), 625–639.
Iskarous, K., Nam, H., & Whalen, D. H. (2010). Perception of articulatory dynamics from acoustic signatures. *The Journal of the Acoustical Society of America*, *127*(6), 3717–3728.
Jacewicz, E., Fujimura, O., & Fox, R. A. (2003). Dynamics in diphthong perception. In *Proceedings of the 15th international congress of phonetic sciences* (pp. 993–996). *Barcelona, Spain*.
Krivokapić, J., & Byrd, D. (2012). Prosodic boundary strength: an articulatory and perceptual study. *Journal of Phonetics*, *40*(3), 430–442.
Lehiste, I., & Peterson, G. E. (1961). Transitions, glides, and diphthongs. *The Journal of the Acoustical Society of America*, *33*(3), 268–277.
Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781.
Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 1–35). Dordrecht: Kluwer.
Lowenstein, J. H., & Nittrouer, S. (2008). Patterns of acquisition of native voice onset time in English-learning children. *The Journal of the Acoustical Society of America*, *124*(2), 1180–1191.
Manuel, Sharon Y. (1990). The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *The Journal of the Acoustical Society of America*, *88*(3), 1286–1298.
Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494.
Moreton, E. (2004). Realization of the English postvocalic [voice] contrast in F1 and F2. *Journal of Phonetics*, *32*(1), 1–33.
Moreton, E., & Thomas, E. R. (2007). Origins of Canadian Raising in voiceless-coda effects: a case study in phonologization. In J. S. Cole, & J. I. Hualde (Eds.), *Laboratory Phonology*, *9* (pp. 37–64). Berlin: Mouton.
Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5), 453–467.
Noiray, A., Iskarous, K., & Whalen, D. H. (2014). Variability in English vowels is comparable in articulation and acoustics. *Laboratory Phonology*, *5*(2), 271–288.
Ohala, J. J. (1997). Aerodynamics of phonology. In *Proceedings of the Fourth Seoul International Conference on Linguistics* (pp. 92–97).
Öhman, S. E. (1966). Coarticulation in VCV utterances: spectrographic measurements. *The Journal of the Acoustical Society of America*, *39*(1), 151–168.
Parker, F., & Walsh, T. (1981). Voicing cues as a function of the tense-lax distinction in vowels. *Journal of Phonetics*, *9*(3), 353–358.
Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, *32*(6), 693–703.
Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-Final consonants in American English. *The Journal of the Acoustical Society of America*, *51*(4B), 1296–1303.
Raphael, L. J., Dorman, M. F., Freeman, F., & Tobin, C. (1975). Vowel and nasal duration as cues to voicing in word-final stop consonants: spectrographic and perceptual studies. *Journal of Speech, Language, and Hearing Research*, *18*(3), 389–400.
Raphael, L. J., Dorman, M. F., & Liberman, A. M. (1980). On defining the vowel duration that cues voicing in final position. *Language and Speech*, *23*(3), 297–307.
Revoile, S., Pickett, J. M., Holden, L. D., & Talkin, D. (1982). Acoustic cues to final stop voicing for impaired-and normal-hearing listeners. *The Journal of the Acoustical Society of America*, *72*(4), 1145–1154.
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.
Saltzman, E., & Kelso, J. A. (1987). Skilled actions: a task-dynamic approach. *Psychological Review*, *94*(1), 84.
Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*(4), 333–382.
Summers, W. V. (1987). Effects of stress and final-consonant voicing on vowel production: articulatory and acoustic analyses. *The Journal of the Acoustical Society of America*, *82*(3), 847–863.
Thomas, E. R. (2000). Spectral differences in /ai/ offsets conditioned by voicing of the following consonant. *Journal of Phonetics*, *28*(1), 1–25.
Thomas, E. R. (2001). *An acoustic analysis of vowel variation in New World English*. Durham, NC: Duke University Press Publication of the American Dialect Society 85.
Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2014). Information for coarticulation: static signal properties or formant dynamics?. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1228–1236.
Walsh, T., & Parker, F. (1984). A review of the vocalic cues to [±voice] in post-vocalic stops in English. *Journal of Phonetics*, *12*(3), 207–218.
Wolf, C. G. (1978). Voicing cues in English final stops. *Journal of Phonetics*, *6*(4), 299–309.