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
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* [bɛ̃nt]) than before voiced ones (as in *bend* [bɛ̃nd]) (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/.

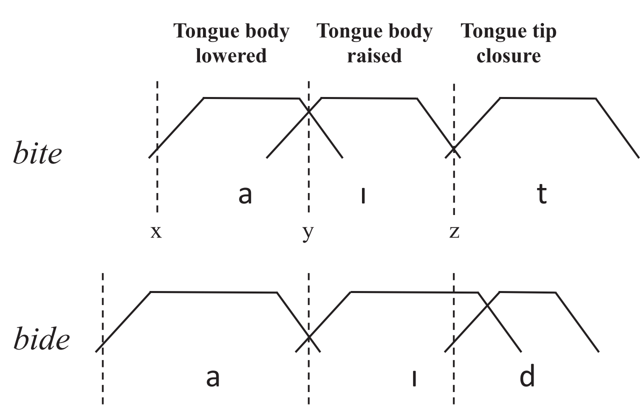

Fig. 1. Proposed gestural scores for words with diphthongal vowels before voiceless codas, as in *bite* [baɪt] (upper panel) and before voiced codas, as in *bide* [baɪd] (lower panel). For clarity, the gestures for the onset consonant are omitted, as is the glottal gesture for the coda. Dashed lines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