Original Paper

Phonetica 2014;71:229–248
DOI: 10.1159/000371820

Received: July 8, 2014
Accepted after revision: December 31, 2014
Published online: April 1, 2015

Gestural Control in the English Past-Tense Suffix: An Articulatory Study Using Real-Time MRI

Adam Lammert[a] Louis Goldstein[b] Vikram Ramanarayanan[a]
Shrikanth Narayanan[a, b]

[a]Signal Analysis and Interpretation Laboratory and [b]Department of Linguistics, University of Southern California, Los Angeles, Calif., USA

Abstract

The English past tense allomorph following a coronal stop (e.g., /bɑndəd/) includes a vocoid that has traditionally been transcribed as a schwa or as a barred i. Previous evidence has suggested that this entity does not involve a specific articulatory gesture of any kind. Rather, its presence may simply result from temporal coordination of the two temporally adjacent coronal gestures, while the interval between those two gestures remains voiced and is acoustically reminiscent of a schwa. The acoustic and articulatory characteristics of this vocoid are reexamined in this work using real-time MRI with synchronized audio which affords complete midsagittal views of the vocal tract. A novel statistical analysis is developed to address the issue of articulatory targetlessness based on previous models that predict articulatory action from segmental context. Results reinforce the idea that this vocoid is different, both acoustically and articulatorily, than lexical schwa, but its targetless nature is not supported. Data suggest that an articulatory target does exist, especially in the pharynx where it is revealed by the new data acquisition methodology. Moreover, substantial articulatory differences are observed between subjects, which highlights both the difficulty in characterizing this entity previously, and the need for further study with additional subjects.

1 Introduction

As is well known, the regular English past-tense suffix takes three phonological forms: /d/ in most contexts: a voiceless allomorph /t/ following voiceless consonants (except /t/), and a syllabic allomorph /Vd/ following coronal (oral) stops. In this lattermost allomorph, the vocoid V has been represented as a short central vowel, such as /ə/ or /ɨ/. While there have been several generative accounts of this allomorphy (Kiparsky, 1985; Pinker and Prince, 1988; Fromkin, 2000; Bakovic, 2005), they have

KARGER 125

0031–8388/15/0714–0229
$39.50/0

E-Mail karger@karger.com
www.karger.com/pho

Adam Lammert
3740 McClintock Avenue, EBB 400
Los Angeles, CA 90089 (USA)
E-Mail lammert@usc.edu

all (implicitly) assumed that the vocoid is a proper vowel segment. This assumption has been made even despite some analyses which identify the vowel as /ɨ/, a contrastive segment that does not occur elsewhere in English, although it can occur as an allophone of the reduced vowel (Flemming and Johnson, 2007). However, by using a grammar in which the coordination of phonological elements in time is represented explicitly (e.g., Browman and Goldstein, 2000; Gafos, 2002; Zsiga, 2003; Davidson, 2006; Goldstein et al., 2006; Nam, 2007), it is possible to hypothesize that such a vocoid emerges from the coordination relation of the consonants on either side of it, rather than from the articulation of a phonological vowel unit. The formalization of this type of 'temporal-only' grammatical specification using coordination constraints can be crucial in developing a satisfactory analysis of alternations, as Gafos (2002) shows for the templatic phonology in Moroccan Arabic. Using such a grammar, the regular English past-tense suffix can be represented as a single coronal gesture whose coordination is determined by phonetic context (Goldstein, 2011). If this analysis is correct, a vocoid can still result from the temporal coordination of consonant units alone, without the need to appeal to a specific spatial target. Thus, it is important to establish whether a spatial target can be identified for this vocoid (which would allow us to reject this analysis), or whether there is no specific spatial target (which would be consistent with the proposed analysis).

Coordination relations have been represented in the grammar by means of a *coupling graph*, that specifies the coupling mode of pairs of planning oscillators that trigger the production of gestural units (Goldstein et al., 2006; Nam, 2007). They have also been represented by markedness constraints that are defined in terms of synchronization of particular landmarks of a pair of units (Gafos, 2002; Zsiga, 2003). In either approach, a targetless vowel interval can emerge from coordination of two consonant gestures (C1 C2) such that C1 is released before the C2 constriction is formed. This will leave a temporal gap between the constrictions whose vocal tract shape is determined by the articulator motions resulting from the release C1 and by movement of individual articulators to their rest postures. Since such a temporal gap will be relatively unconstricted and variable (its vocal tract shape will largely be determined by flanking consonants and vowels), this seems a plausible model for reduced vowels, such as the one that appears in the syllabic past-tense allomorph. Browman and Goldstein (1992) tested the hypothesis that all schwa vowels in English were 'targetless' in this sense, but they rejected it. There was evidence for a specific constriction target associated with those schwa vowels. However, their materials were nonsense forms and did not examine any affixes. The vowel of the plural affix (e.g., 'roses') has been shown (Flemming and Johnson, 2007) to be acoustically different from the stem schwa vowels in final syllables (e.g., 'Rosa's'), in a way that is consistent with the traditional transcription of the plural (and past tense) affixes as [ɨ], but the stem (or 'lexical', as it will be referred to here) vowel as [ə] (Trager and Smith, 1951). Flemming and Johnson (2007) showed that the plural affix vowel is higher (has a lower F1) than the one found in final lexical schwas, suggesting that the plural vowel is more constricted. Since F2 in these affix vowels ranges from 1,750 to 2,200 Hz, consistent with a fronted tongue shape, the acoustics seem consistent with a tongue body position for a coronal consonant. Thus, these reduced vowels appear to be good candidates for purely temporal gaps between release of one coronal and formation of the next. Since similar transcriptions have been given for the past-tense suffix, these appear to be possible candidates for the past-tense allomorph as well, as hypothesized in Goldstein's (2011) analysis.

The targetless hypothesis for the past-tense affix was tested by Smorodinsky (2002), who used EMA to collect kinematic data from the tongue, jaw, and lips while speakers read utterances like those in (a), with near-minimal pairs of lexical versus past-tense affix vowels:

(a) 'If Cheetah'd even known' (lexical)
'If cheated even once' (affix)

That work found that the position of the tongue receivers during the affix vowels were more correlated with the positions of the flanking full vowels than was the case for the lexical vowels, which was taken as evidence that the affix vowels added no control of their own to the shaping of the vocal tract. However, while EMA provides data with a relatively clean audio signal and excellent temporal resolution, it suffers from poor spatial resolution in the sense that the tracked flesh points are few and widely spaced, and the entire midsagittal view of the vocal tract is not imaged. The latter two properties are serious shortcomings when studying vowels and other vocoids, where shaping along the entire length of the vocal tract should be considered. Thus, in the current study, we make use of real-time magnetic resonance imaging (rtMRI) (Narayanan et al., 2004; Bresch et al., 2006), which trades some temporal resolution for improved spatial resolution as well as complete midsagittal views of the vocal tract, including pharyngeal regions. We collected midsagittal images from 2 male, native English speakers. The speakers read phrases aloud which contained either lexical schwas or past-tense affix reduced vocoids in multiple contexts. For convenience, we refer to these reduced vowels as lexical versus affix schwas (with the understanding that the affix schwa is often transcribed as a barred i, not actually schwa). Images of the different schwas and their context were extracted from the rtMRI data and analyzed.

The goals of this study are both empirical and methodological. Our primary goal is to test predictions consistent with the hypothesis of a targetless affix schwa. To that end, we characterize and compare lexical and affix schwas in terms of posture, duration, acoustics and articulatory targets. Regarding the lattermost point, a general statistical method is developed for evaluating targetedness of articulatory events within the framework previously presented by Browman and Goldstein (1992) in their effort to quantify targeted versus targetless articulations based on their predictability from context.

We also take a novel approach to processing the particularly rich data afforded by rtMRI. The first step in a conventional approach would be to infer air-tissue boundaries in the images (e.g., Bresch and Narayanan, 2009), which is anatomically well founded, but tends to have high computational cost and lacks robustness. We circumvent these problems by directly examining the pixel intensities. Given a location in the midsagittal plane, the pixel intensity values will vary across a collection of images as a direct result of variation in the vocal tract configurations represented. The usefulness of related methods has recently been demonstrated on rtMRI data to extract vocal tract constriction information robustly and efficiently (Bresch et al., 2010; Lammert et al., 2010).

Section 2 of this article describes our methodology, including data collection and processing, as well as the development of our statistical hypothesis tests. In section 3 we present the results of our analysis, which are then discussed in section 4. Finally, in section 5 we present our concluding remarks.

2 Method

2.1 Data Acquisition

Data were collected on October 26th, 2008 at the University of Southern California Health Sciences Campus in Los Angeles, Calif. The subjects (denoted M.B. and B.P.) were 2 white males in their mid-20s, who were enrolled as graduate students at the University of Southern California at the time of participation. Both subjects were native speakers of American English, raised in the northern United States (Rochester, N.Y. and Madison, Wisc., respectively). Neither subject reported any history of speech or hearing pathologies. Both subjects were naïve to the purpose of the study. All work was approved by the University of Southern California institutional review board prior to acquisition.

Acquisition of MR images was performed on a GE Signa™ 1.5-tesla scanner. The rtMRI protocol implemented a 13 interleaf spiral gradient echo pulse sequence with a repetition time of TR = 6.5 ms. The field of view covered an area of 20 × 20 cm oriented along the midsagittal plane. Slice thickness was approximately 6 mm. Images were reconstructed using a sliding-window procedure with a step size of 1 TR, resulting in an effective frame rate of 153.85 fps.[1] The spatial resolution of the reconstructed images was 68 × 68 pixels (approximately 2.9 mm pixel width). Further details regarding our rtMRI protocol can be found in Narayanan et al. (2004). Synchronous audio recordings of the subjects' speech were also acquired using an optical microphone and were subsequently denoised according to the protocol described in Bresch et al. (2006). The denoising technique promises nearly 30 dB noise suppression during speech and, in general, minimal errors are expected in terms of spectral distortion as a result of applying this method. However, one may expect, due to the nature of the recording environment inside the scanner bore, some reverberation and background noise caused by the cryogen pump and ventilation system.

Stimuli closely followed those from the previous study by Smorodinsky (2002), in that they consisted of phrases with embedded schwas. All phrases contained the sequence $/v_1 c_1 v_p c_2 v_2/$, where $v_1 = v_2 \in$ {i, ɪ, eɪ, ɛ, æ, ɑ, oʊ, u}, $p \in$ {lexical, affix} and $c_1 = c_2 =$ /d/. This allowed a total of 16 unique phrases (= 8 vowels × 2 schwa types). Seven sets of these phrases were created by randomizing the order of all 16 phrases 7 times, for a total of 112 phrases, 56 lexical schwas and 56 affix schwas[2]. All seven sets of stimuli can be seen in Appendix A.

Subjects were instructed to lie comfortably inside the scanner in supine position while they read the stimuli from paper sheets suspended in front of them at a distance of approximately 15 cm. Subjects had their heads padded into place near the temples to prevent lateral motion. Paper sheets each contained half of one set (i.e., eight phrases). These half-sets could be read within 30 s (approximately 3 s per phrase), which is the limit for running the specified protocol without risk of overheating the scanner. Subjects were allowed to repeat any stimuli if they believed their production to be in error. For the purposes of analysis, errorful trials were eliminated from consideration and only the final production of the given token was used. Additional repetitions were recorded in separate scans, immediately following the scan in which they were originally spoken.

A potential challenge for interpreting any speech data collected from rtMRI is the supine position assumed by subjects during acquisition (e.g., Wrench et al., 2011). While some studies have found notable differences in tongue posture during the production of sustained speech sounds (Kitamura et al., 2005), other studies have consistently revealed very small differences during continuous speech (Tiede et al., 2000; Stone et al., 2007). Because all speech analyzed in this study was continuous in nature, small differences in posture should therefore be expected. Additional evidence suggests that articulations which are critical to the production of a given phoneme are relatively less influenced of gravity. Since our methodology is, in one sense, intended to uncover those articulations which are critical (i.e., targeted), we should be isolating precisely those aspects of articulation which are least impacted by the effect by the supine posture. Moreover, since all images were acquired with subjects in supine position, there is no possibility for more subtle confounds due to cross-posture comparison.

[1] Note that, using 1-TR reconstruction, images spaced less than 13 TR = 84.5 ms apart will be composed of some overlapping frequency-domain data.

[2] A single lexical schwa from M.B. was later eliminated from further analysis due to an error at acquisition time.

2.2 Processing

Representative points for each segment in the /$v_1c_1v_pc_2v_2$/ sequence were annotated by combined inspection of formant and intensity analyses using Praat (Boersma, 2001). Local minima in intensity surrounding the schwa (t_{c_1} and t_{c_2}) were identified, corresponding to the flap consonants. Time points immediately preceding the formant transition into the first coronal flap (t_{v_1}) and immediately following the formant transition out of the second flap (t_{v_2}) were identified and used as points corresponding to the context vowels. These points were chosen because they represent the closest stable, representative vowel articulations to the schwa, and also because choosing the temporal center of the vowel is unsuitable for vowels that are diphthongized, such as oʊ and eɪ. The schwa vocoid was taken to be the point exactly between the coronal closures [i.e., $t_{v_p} = (t_{c_1} + t_{c_2})/2$]. At these annotated points, formant frequencies were extracted and MR images were reconstructed at a spatial resolution of 68 × 68 pixels (2.7 mm/pixel). Given the MRI protocol described above, 13 frequency-domain samples are required to provide complete coverage of the spatial frequency domain and reconstruct a single image. In this case, we took the temporally closest sample to the annotated points, along with the 6 samples before and after that sample.

An rtMRI video sequence may be regarded as a spatially and temporally varying function of gray-level intensity values. This function can be denoted as I [*m*, *n*, *t*], where *m* and *n* represent the vertical and horizontal position of a pixel in the image plane, respectively, and *t* is the time associated with a particular video frame. In the analysis presented here, the intensity of each pixel in the image plane is treated as an articulatory feature. Given an image 68 × 68 pixels in size, this provides a total of 68 × 68 = 4,624 features per image. Retaining such a large number of features may seem unwieldy, but it provides the opportunity to explore information about the entire midsagittal plane, while making minimal assumptions about what information might be important to articulation. Pixelwise analysis is also relatively robust in comparison to the more traditional approach, involving edge detection and extraction of air-tissue boundaries (see Lammert et al., 2010) for a detailed discussion). Of course, many of these articulatory features are irrelevant from the perspective of speech production (e.g., locations anterior to the face and superior to the nasal cavity). Thus, not all pixels in the image plane were used as articulatory features. Pixels considered to be articulatory features were limited based on the observed variation in pixel intensities across all extracted images, as measured by the standard deviation. Only those pixels which had standard deviation values at or above the 80th percentile of standard deviations across all pixels – a total of 925 pixels – were considered in the analysis of targetedness described below.

Even though subjects had their heads padded in place, small amounts of head motion were sometimes observed between scans (i.e., from one half-set to the next) if subjects shifted their body position slightly for comfort. This motion must be corrected because the pixelwise analysis utilized here rests on the assumption that the image plane and the midsagittal plane coincide. Toward correcting for head motion, it is assumed that head motion happens over a longer time scale than a single phrase, which is reasonable because the only observed motion was between sets of eight phrases. It is further assumed that head motion is accurately represented as a rigid transformation of the head. While it is possible that some head motion may be nonrigid in nature – for instance, due to stability of the spine and subglaryngeal vocal tract as the head moves – a rigid transformation is an appropriate approximation in this case because the primary view is of the supralaryngeal vocal tract and because the amount of head motion is small.

Given these assumptions regarding head motion, a simple brute-force algorithm can be applied to find the optimal rigid transformation of the head for all images in a phrase (Forsyth and Ponce, 2002). This can be done by comparing a representative vowel image for each phrase against some overall template image of the vowel in that phrase. The template image for a given vowel context was defined as the mean image of all recorded context vowels of that type. The representative image for a particular phrase was taken to be the mean image of both context vowels in that sentence. For a specific phrase, the algorithm proceeds to calculate the normalized two-dimensional cross-correlation between the template image and a series of rotated versions of the representative image. The rotation and cross-correlation offset corresponding to the maximum cross-correlation value is taken to be the optimal translation and rotation. The optimal rotation and translation obtained from this procedure were then implemented in an affine transformation matrix which associates Cartesian coordinates in the representative image plane to those in the template image plane. The transformation specified by this matrix was then applied, with bilinear interpolation, to all images of interest within that phrase

Table 1. Mean and standard deviation (SD) of rigid transformation parameters (translation and rotation) applied to all images as head motion correction for each subject.

	M.B.		B.P.	
	mean	SD	mean	SD
θ, °	–0.01	0.07	–0.03	0.62
tv, mm	–0.02	0.44	0.53	1.63
th, mm	0.02	0.26	0.58	1.61

Both subjects exhibited very little head motion overall. Subject M.B.'s head remained very stationary across all tokens, while subject B.P. displayed slightly more motion. Degrees are defined in the clockwise direction.

(i.e., vowels, closures and schwa). Using this algorithm, the amount of head motion can also be quantified. The results of this quantification can be seen in table 1. The overall amount of head motion was quite small, with subject B.P. displaying slightly more than M.B. The increment of translation was equal to one pixel width, and the increment of rotation was equal to 0.25°. The algorithm described here was implemented in Matlab® (The MathWorks Inc., version 7.8.0) using the Image Processing Toolbox™.

2.3 Articulatory Target Analysis

A method was developed for testing whether observed articulatory kinematics can be considered targeted or not, which aims at the central focus of the present investigation. The method expands upon the analysis of targetedness presented by Browman and Goldstein (1992) and adopts the assumptions of that analysis. Most importantly, a definition of targeted versus targetless articulation was adopted in the present study, which is based on the level of predictability of a given articulation from the surrounding segmental context. In particular, a targeted articulation cannot be entirely explained without attributing some aspects of the articulation to factors outside of the context, namely the target itself. On the other hand, a targetless articulation can be defined as entirely predictable from contextual articulation. Previous studies have shown the power of contextual segments in predicting schwa-related production behavior, including those directly adjacent to the schwa, and also the nearest vowels whether they are adjacent or not (Anderson, 1982; Magen, 1989).

Note that this definition of targetlessness is based entirely on coarticulatory effects, and it assumes that those effects are directed only from the surrounding context toward the segment of interest. It is also possible, however, that these effects could show influence in the reverse direction. If an articulation that would be judged as targeted by some other criterion exerts strong coarticulatory influence on its context, it might be judged as targetless by the present definition. This issue does not affect the interpretation of articulations that are judged to be targeted by the presently proposed method, therefore the present focus will be on developing an analysis to identify targeted behavior. However, any behavior that is determined to be not targeted by the proposed methods will be called 'targetless', partially as a term of convenience.

As stated more precisely by Browman and Goldstein (1992), the issue of targeted versus targetless articulation can be defined as a choice between two competing models of the articulatory variable in question. Elements of these competing models are articulatory features representing the relevant segmental context which, based on studies mentioned above, is taken to include v_1, c_1, c_2 and v_2. Models developed for other articulations of interest may contain different definitions of what constitutes the relevant context (one alternative definition will be explored below in validating the presently proposed method). Remembering that the articulatory variables being considered here are the intensities of individual pixels in the image plane, one can assume a linear model and express the articulation of interest, $y = \mathrm{I}\,[m, n, t_{v_p}]$, as a weighted combination of the surrounding context:

$$\hat{y}_{targetless} = b_{v_1}\,\mathrm{I}\,[m, n, t_{v_1}] + b_{c_1}\,\mathrm{I}\,[m, n, t_{c_1}] + b_{c_2}\,\mathrm{I}\,[m, n, t_{c_2}] + b_{v_2}\,\mathrm{I}\,[m, n, t_{v_2}] \quad (1)$$

This is the basic model for a targetless entity, since there is no term in the model corresponding separately to the entity itself. All terms relate to the segmental context. If the articulation being predicted is targeted, an additional term can be added to the model, leading to an equation of the following form:

$$\hat{y}_{targeted} = b_{v_1} \mathrm{I}\,[m, n, t_{v_1}] + b_{c_1} \mathrm{I}\,[m, n, t_{c_1}] + b_{c_2} \mathrm{I}\,[m, n, t_{c_2}] + b_{v_2} \mathrm{I}\,[m, n, t_{v_2}] + k \quad (2)$$

Note that the two models differ only by the presence or absence of a constant offset term, k, representing a separate articulatory target specifically for the entity in question. Fitting both models (i.e., estimating the values of the parameters b and k) was treated as a standard multiple linear regression problem, and solutions were found in the traditional least-squares sense.

It is worth stating explicitly that these models are extremely simple in terms of the articulatory information they take into account. They do conform logically to the definition of targetlessness and targetedness mentioned above and, moreover, it will be shown here that they provide a good fit to the data despite their simplicity. Still, neither model takes into account any higher-order polynomial dependencies on the context. It should also be noted that the models, as written, do not take into account the time between articulatory events, nor do they consider the articulatory trajectories taken from one posture to another. Both state a simple linear relationship between a specific schwa-related articulatory posture and representative articulatory postures from the context.

The problem of determining targetedness can then be treated as a problem of model selection – that is, choosing which model is more appropriate for explaining the observed data. Appropriateness of a given model can be considered in terms of its accuracy on a particular data set (e.g., absolute residual error), but also in terms of the ability of the model to generalize to new data sets. Both of these criteria can be assessed in a principled way using cross-validation (Devijver and Kittler, 1982; Geisser, 1993), which is a method for estimating, from how a model fits the current sample, how that same model will fit the entire population. It works by iteratively partitioning a data set into a training and test set, which are used for fitting and evaluating the model, respectively. At each iteration, the accuracy of the model on the test set is compiled. For the present purposes, four separate cross-validations were performed for each subject: twice for each model (i.e., targeted vs. targetless) and twice for each schwa type. The number of data partitions made was equal to the number of data points, which is often known as a leave-one-out scheme.

After all iterations are complete, the overall performance of the model can be compared to the performance of other models for the purposes of selection. In most situations, it is appropriate to simply select the model which produces the smallest overall residual error, because this model will represent a compromise between under- and overfitting the data. However, in the current setting, the data set is large and the models are simple, making it very unlikely that we will ever observe overfitting of the data. Consider, for instance, that there are approximately 56 examples of each schwa in our data, but only four or five model parameters to estimate. Therefore, there will be between 11 and 14 examples per parameter, likely making this situation highly overdetermined. For the reasons stated above, a statistical hypothesis test was developed to select between models 1 and 2. At each iteration of the cross-validation procedure, the absolute residual error of the competing models is calculated:

$$\rho = |\hat{y}_{targetless} - y| - |\hat{y}_{targeted} - y| \quad (3)$$

This quantity, ρ, is a statistic that will be positive when the absolute residual error of the targeted model (eq. 2) is lower than that of the targetless model (eq. 1). The distribution of ρ across all cross-validation iterations can then be used as an indication of the appropriateness of selecting the targeted model over the targetless model. If the articulation of interest, y, is truly untargeted, the absolute residual error resulting from cross-validation should be approximately equal for both models, and the distribution of ρ should have a mean of zero. On the other hand, if the entity of interest is truly targeted, the absolute residual error of the targetless model should be significantly higher because it will produce a poorer fit to the data and the distribution of ρ should have a mean significantly higher than zero.

Whether the distribution of ρ is mean zero or significantly higher can be tested with a right-tailed Student's t test (Appendix B provides justification for using a parametric test in this instance) with the null hypothesis that the mean of ρ across all cross-validation iterations is zero. A rejection of the null hypothesis at the $\alpha = 0.05$ level is interpreted as high confidence that the targeted model is more

appropriate. This significance threshold has been the subject of much debate over the increase in type I errors that can result from performing multiple comparisons without adjusting the familywise error rate (Bennett et al., 2010). Lowering the significance threshold, on the other hand, is associated with a variety of complementary problems, including increases in type II errors, difficulty in delimiting the family within which to adjust the threshold and several other conceptual problems (Rothman, 1990; Perneger, 1998; Feise, 2002). Since the primary objective of this hypothesis test was essentially exploratory and not confirmatory (i.e., to 'discover' which regions of the image are targeted), a significance threshold was chosen which reflects higher concern about type II errors than type I. However, clusters of significant pixels of size less than two, as defined using 8-neighbor connectivity, were discarded under the assumption that targeted behavior should occur over areas larger than a few millimeters and that any such significant tests were spurious. These eliminations increase the effective significance threshold, but do so in a way that leverages the knowledge that spatially adjacent pixels are likely to be correlated over time.

After the model selection procedure, the cross-validation scheme is put aside and the selected model is fit one final time on all examples of the articulation in question. The parameters resulting from this final fit can then be subjected to the usual interpretations assigned to regression models. Of particular importance here is the intercept term estimated for those articulatory features (i.e., pixels) that are determined to display targeted behavior. The intercept term can be interpreted as the target itself, and it is possible to infer the magnitude and direction of articulatory movement associated with the articulatory target by examination of its numerical value. The sign of this term gives an indication of the movement of articulator tissue suggested by the target. Positive values of the target term indicate a preference for the movement of tissue (e.g., the tongue) into the location of a pixel during production of the schwa, relative to the context.

This test for targetedness was validated by assessing the targetedness of the coronal closures, c_1, in our data set. Coronal closures were chosen because a clear articulatory target can be assumed for those segments a priori, namely contact between the tongue blade and/or tip and the palate. At the same time, other aspects of articulation, such as lip aperture, should be highly predictable from the surrounding vowel context. Following from the models presented in equation 1 and 2, we define two similar models that utilize the vowels surrounding the closure – v_1 and v_p – as the relevant segmental context. Specifically, the competing models are as follows:

$$\hat{y}_{targetless} = b_{v_1} \mathrm{I}\,[m, n, t_{v_1}] + b_{v_p} \mathrm{I}\,[m, n, t_{v_p}] \tag{4}$$

and

$$\hat{y}_{targeted} = b_{v_1} \mathrm{I}\,[m, n, t_{v_1}] + b_{v_p} \mathrm{I}\,[m, n, t_{v_p}] + k \tag{5}$$

Aside from adjustments to the models, the cross-validation procedure and hypothesis testing was carried out as described above.

The results of this analysis can be seen in figure 1. The vocal tract images in this figure show the overall mean posture across all recorded examples c_1, calculated as the pixelwise mean value of $\mathrm{I}\,[m, n, t_{c_1}]$ for all examples of c_1, and repeated for all pixel locations (m, n) in order to form an image. Pixels that were determined to display targeted behavior are highlighted with superimposed circles. The colors of the circles corresponds to the value of the offset term (i.e., the articulatory target) estimated in the final fitting of the model, after the model selection procedure is completed. It is clear from the concentration of highlighted pixels in the alveolar region of both subjects that constriction in that region is a highly targeted activity, as expected. The fact that the target is, indeed, a constriction action is evidenced by the values of the estimated offset terms, indicating a strong preference for tissue moving into the alveolar region during the consonant. Note that the tongue tip of subject B.P. is not highlighted as targeted. It was observed in the images that B.P. (a) often presents a /d/ that is highly flapped and does not display complete close in the midsagittal plane, and (b) the precise position of B.P.'s tongue tip is highly variable in a range of approximately 1 cm along the alveolar ridge. Both of these may lead our test for targeted behavior to conclude that specific placement of the exact tongue tip is not part of the target for this token. Subject M.B. shows additional targeted activity under the tongue blade characterized by tissue moving out of that region, which can be taken as evidence of sublingual cavity formation during the production of this coronal consonant. The targeted pixels along

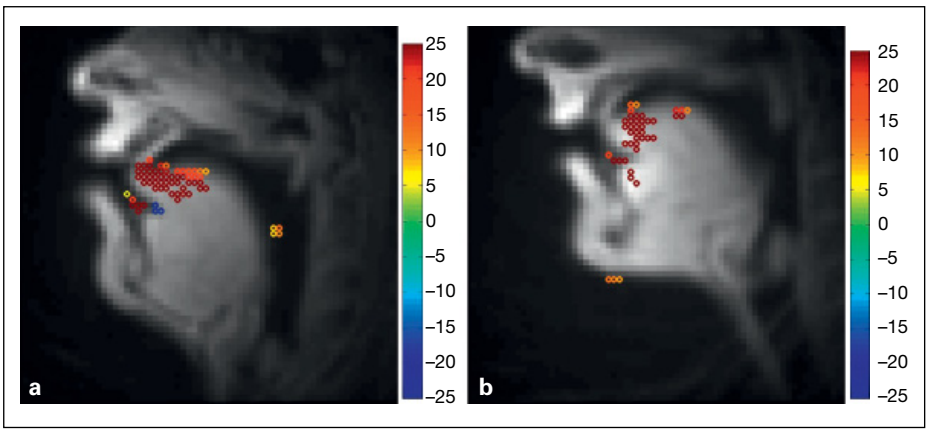

Fig. 1. Results of applying the test for targetedness to the coronal closure, c_1, for the purposes of validating the test. Vocal tract images show the overall mean posture across all examples of c_1. Targeted pixels are highlighted with superimposed circles. The colors of the circles represent the value of the offset term (i.e., the articulatory target). As should be expected, strong targeted behavior is seen in the alveolar region, and the value of the offset term indicates articulatory movement into that region during the consonant, or constriction action. **a** M.B., lexical, red = targeted pixels. **b** B.P., affix target direction.

the jaw for subject B.P., as well as the pixels in the pharynx for subject M.B., are assumed to represent secondary articulations that are speaker-specific for this token. Moreover, the general lack of targeted activity near the lips and velum is consistent with expectations.

2.4 Additional Analysis

Differences between lexical and affix schwas were analyzed along several additional descriptive dimensions, both articulatory and acoustic, in addition to the hypothesis test for targeted aspects of the articulation in the vocal tract. This additional information is essential in providing an interpretation of the differences in production between lexical and affix schwa. Articulatory differences were analyzed by examining the overall mean posture during the vocoids in question. Acoustic differences were analyzed according to acoustic duration, as well as distribution in the (F1-F2) formant frequency plane.

Duration differences between lexical and affix schwas were examined using the acoustic annotations previously described. In particular, schwa duration was defined as the interval between the temporal center of the preceding coronal closure until the temporal center of the following coronal closure ($t_{c_2} - t_{c_1}$). The mean and standard deviation of duration were calculated across all tokens of each schwa type, and histograms were calculated using 10 equally spaced bins between 50 and 200 ms to compare the distribution of duration.

Acoustic analysis of lexical and affix schwas was performed by examination of the first and second formant frequencies (F1 and F2) measured at time t_{v_p}. Formant tracking analysis was performed using Praat (Boersma, 2001). Formant tracking was configured to find five formants in the range from 0 to 4,000 Hz with a window length of 25 ms. These parameters were found to be optimal by manual tuning and visual inspection of the estimated formants overlaid on spectrograms of the spoken utterances. Descriptive statistics, including the mean, standard deviation and covariance were calculated over all instances of both schwa types in the F1-F2 plane. The mean positions of the full context vowels in the F1-F2 plane were also calculated at time t_{v_1} and t_{v_2} to provide an anchor for examining the distribution of lexical and affix schwa.

Postural differences in vocal tract configuration were examined between lexical and affix schwa by examining the overall mean vocal tract posture as calculated across all recorded tokens for both vocoids. This was done by taking the pixelwise mean value across all tokens associated with

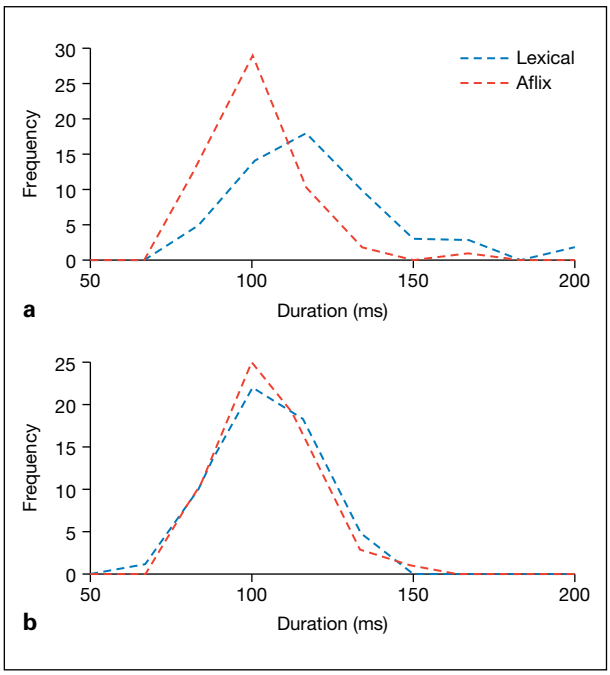

Fig. 2. Distribution of lexical and affix schwa durations for both subjects. Subject M.B. (**a**) displays a much longer and more variable lexical schwa as compared to affix schwa. Subject B.P. (**b**) shows nearly identical distribution of duration for both schwa types.

a particular schwa type and speaker. In other words, the mean of I [m, n, t_{v_p}] was calculated for all examples of v_p corresponding to a particular schwa type and a particular speaker, and this mean calculation was repeated for all pixel locations (m, n). This collection of pixelwise mean values is itself an image, representing the average vocal tract posture of the schwa in question. The pixelwise difference between mean images – which can also be regarded as an image – was also calculated to facilitate comparison of vocal tract configurations. Difference images were calculated by subtracting the lexical schwa mean image from the affix schwa mean image for both subjects.

3 Results

Figure 2 shows the distribution of schwa durations for both subjects. The distribution of lexical schwa durations for subject M.B. (mean = 121 ms, SD = 25 ms) is noticeably shifted toward longer times, as compared to affix schwa durations (mean = 102 ms, SD = 16 ms). Subject B.P. displayed much more consistency in duration of lexical schwa (mean = 105 ms, SD = 15 ms) and affix schwa (mean = 104 ms, SD = 14 ms). Two-sample t tests were performed for both subjects to test the hypothesis that the samples of lexical and affix schwa were drawn from populations with equal mean durations ($\alpha = 0.05$) against the hypothesis that they were not. Results show that the mean duration is significantly different for subject M.B. [$t(109) = 5.003$, $p << 0.05$], but not significantly different for subject B.P. [$t(110) = 0.107$, $p = 0.92$].

Figure 3 shows the mean positions of the full context vowels in the F1-F2 plane, and the distribution of lexical and affix schwas is indicated for each subject by an overlaid ellipse. Ellipses correspond to one standard deviation from the mean, as determined by calculation of the full covariance matrix between F1 and F2. The distribution

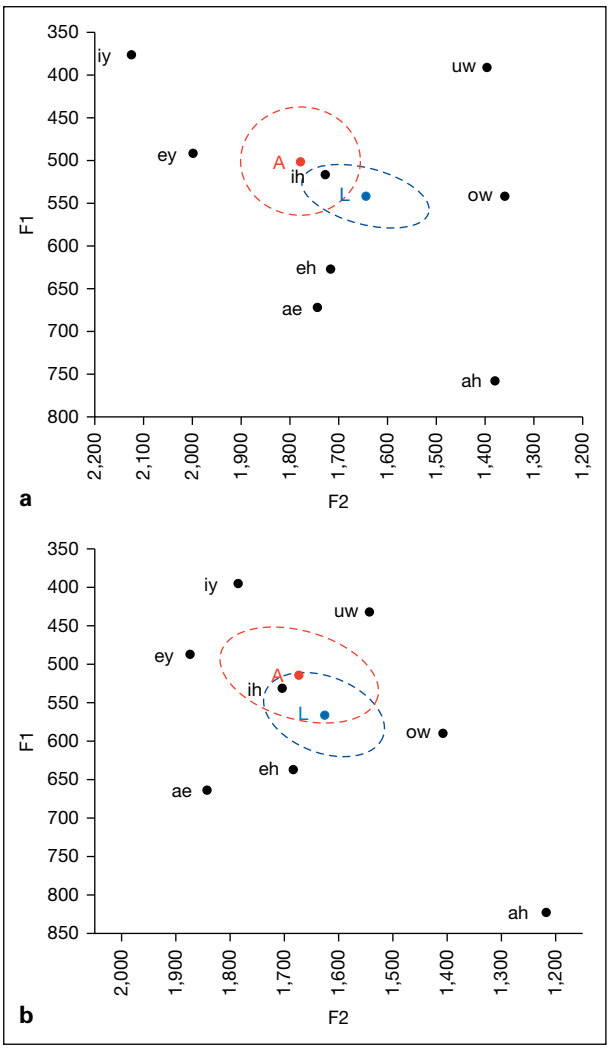

Fig. 3. Results of formant frequency analysis, showing the distribution of mean full vowel position in the F1-F2 plane for subjects M.B. (**a**) and B.P. (**b**). Values of F1 and F2 are in Hertz. Vowel labels are given in Arpabet. Overlaid are the mean position of lexical (blue, labeled L) and affix (red, labeled A) schwas along with two ellipses representing one standard deviation from the mean.

of lexical schwas is centered slightly lower and further back than /ɪ/ and more broadly distributed in the direction defined by /eɪ/–/oʊ/, as appears to be the case for both subject M.B. (F1: mean = 542 Hz, SD = 37 Hz; F2: mean = 1,645 Hz, SD = 128 Hz) and subject B.P. (F1: mean = 567 Hz, SD = 53 Hz; F2: mean = 1,626 Hz, SD = 111 Hz). The distribution of affix schwas for subject M.B. is centered both higher and further forward than for lexical schwas and is considerably broader in the F1 dimension (F1: mean = 502 Hz, SD = 63 Hz; F2: mean = 1,778 Hz, SD = 120 Hz). The distribution of affix schwas for subject B.P. is centered slightly higher than for lexical schwas (F1: mean = 514 Hz, SD = 61 Hz; F2 mean = 1,673 Hz, SD = 143 Hz) but retains a similar but slightly broader pattern of variance. In the spirit of analyses presented by Flemming and Johnson (2007), two-sample t tests were performed for both subjects to test the hypothesis that the samples of lexical and affix schwa were drawn from

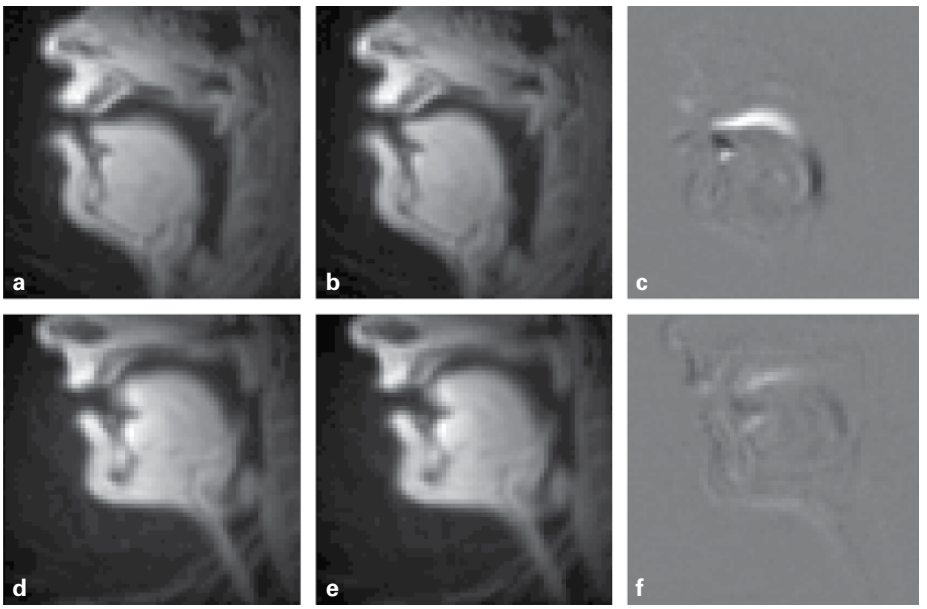

Fig. 4. Vocal tract images showing the overall mean posture across all examples of each schwa type, for both subjects, as well as the difference images (**c**, **f**) which highlight the differences between schwa types. Both subjects display raised tongue position in the affix schwa case, although the effect is much more dramatic for subject M.B. **a** M.B., lexical. **b** M.B., affix. **c** M.B., affix – lexical. **d** B.P., lexical. **e** B.P., affix. **f** B.P., affix – lexical.

populations with equal mean F1 ($\alpha = 0.05$) against the hypothesis that they were not. Similar t tests were performed for F2, as well. Data from the 2 subjects were not pooled in order to assess their behavior independently, which also eliminated the possibility of performing a single two-way ANOVA for each formant. Results for subject M.B. show that mean F1 for the two schwas is significantly different [$t(109) = 4.02$, $p << 0.05$]), as is mean F2 [$t(109) = -5.62$, $p << 0.05$]. Results for subject B.P. show that mean F1 for the two schwas is significantly different [$t(110) = 4.76$, $p << 0.05$], whereas mean F2 is slightly above the significance threshold [$t(110) = -1.94$, $p = 0.055$].

Figure 4 displays the mean vocal tract posture for each subject producing both lexical and affix schwas. Although the overall vocal tract configurations appear similar, differences in tongue height can be observed for both subjects between schwa types. These differences are highlighted by the difference images, which show regions of the midsagittal plane where the articulators are present (indicated by bright pixels) or absent (dark pixels) during production of an affix schwa versus a lexical schwa. Specifically, affix schwas display a higher tongue position as compared to lexical schwas. This effect is much more dramatic for subject M.B., but also present for subject B.P.

Advancement of the tongue dorsum in the uvular-upper pharyngeal region can also be observed for subject M.B. during affix schwa.

Figure 5 shows the results of applying our test for targetedness to the data from both subjects and both schwa types. Both subjects M.B. and B.P. display the largest

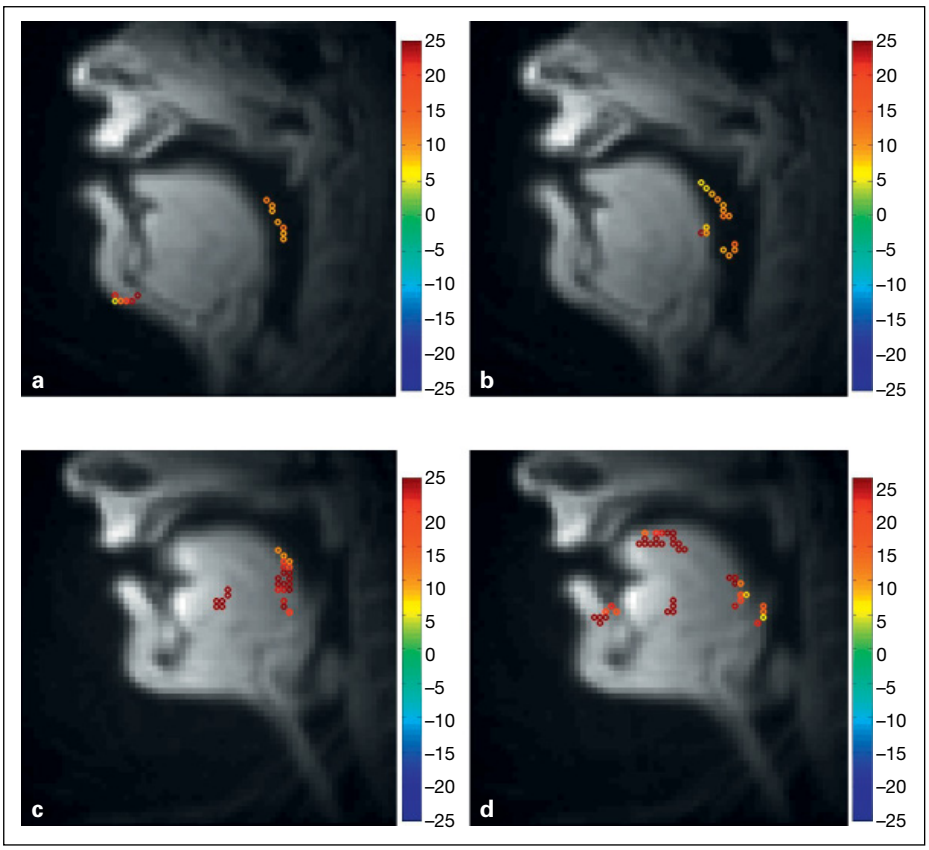

Fig. 5. Results of applying the test for targetedness to examples of lexical and affix schwa. Vocal tract images show the overall mean posture across all examples of each schwa type. Targeted pixels are highlighted with superimposed circles. The colors of the circles represent the value of the offset term (i.e., the articulatory target). During lexical schwa production, subjects M.B. (**a**) and B.P. (**c**) display targeted constriction behavior in the midpharynx, and subject M.B. shows lowering of the jaw. During affix schwa production, M.B. (**b**) shows targeted constriction behavior in the lower and upper pharynx, while B.P. (**d**) displays a lower-pharyngeal constriction target and a substantial constriction target in the palatal region.

concentration of targeted pixels in the middle pharynx for lexical schwa, with M.B. also displaying a concentration at the jaw, indicating jaw lowering. For affix schwas, both M.B. and B.P. show targeted areas in the lower pharynx. In addition to this, M.B. shows targeted areas in the upper pharynx and B.P. shows a large concentration of targeted areas in the palatal region, indicating a substantial constriction target in that region, likely achieved by increased tongue height.

Table 2 shows the mean RMS error of the targetless model (eq. 1) and the targeted model (eq. 2) across all pixels that were determined to be targeted. These measurements provide a quantification of the goodness of fit in terms of each model's ability to predict the pixelwise articulatory features. Higher values indicate more error, and therefore poorer fit. Results are divided by speaker, schwa type and model type. The

Table 2. RMS error of the targetless model (eq. 1) and the targeted model (eq. 2) across all pixels that were determined to be targeted for both speakers

	M.B.		B.P.	
	lexical	affix	lexical	affix
Targetless model (eq. 1)	3.89	3.26	6.78	7.75
Targeted model (eq. 2)	3.32	2.81	5.81	6.54

Units are gray-level pixel intensity values, which have a nominal dynamic range of 0–255. These values provide a quantification of the overall goodness of fit of the models, with higher values indicating poorer fit. Modeling errors were generally low, indicating that both models are capable of modeling the data well. The targeted model provided a better fit, in general, which is expected from the addition of more parameters.

units of these error values are gray-level pixel intensity, which have a nominal dynamic range of 0–255. All error values were less than or approximately equal to 3% of this range, which indicates that both models are capable of providing a quite accurate fit to the data. In addition, the low errors indicate that the signal-to-noise ratio of the images was favorable. The targeted model provided, in general, a better fit than the targetless model, which is expected from the addition of one more parameter in the targeted model. The goodness of fit was better on subject M.B.'s data versus B.P.'s data. This difference in fit between subjects is not likely to be informative, given that the model fits were so good overall, and may be due to small differences in image noise, speaking style, or even craniofacial morphology.

4 Discussion

Lexical schwa can be well characterized in both acoustic and articulatory terms. Acoustically, the distribution of lexical schwa in the F1-F2 plane is highly similar between subjects, situated slightly below and back of /ɪ/, with a relatively tight distribution over that acoustic region. In terms of articulation, both subjects display targeted movement of the tongue back into the midpharyngeal region during production of lexical schwa. Subject M.B. also displays targeted behavior with respect to lowering of the jaw during lexical schwa, which likely aids in moving the tongue down and back. Because the jaw is massive and slow-moving, this jaw target may be directly related to the long duration of lexical schwa displayed by M.B., but it is not clear which is the causal factor – that is, whether the need to employ the jaw slows down production, or whether slower production affords the time required to employ the jaw. Subject B.P. displays neither the slower production nor the jaw-related target.

Affix schwa is consistently different from lexical schwa, both in terms of articulation and acoustics. It is more acoustically variable than lexical schwa and consequently less centralized overall. Affix schwa is also generally higher than lexical schwa and – especially for subject M.B. – further forward. This acoustic characterization is consistent with the higher average tongue postures observed for both subjects during affix schwa as compared to lexical schwa (fig. 4), which in turn is consistent with previous

characterizations of affix schwa as /ɨ/ (e.g., Flemming, 2007). However, this characterization may not be precisely appropriate across both subjects, since subjects seem to show dramatic differences in how much higher the tongue is during affix schwa versus lexical schwa, and since the pattern of acoustic variability is quite different between subjects.

The observation that affix schwa is more acoustically variable than lexical schwa is also consistent with the hypothesis that affix schwa is a targetless vocoid. However, the articulatory analysis presented here strongly suggests an articulatory target for affix schwa in which the lower pharynx is constricted relative to its posture during the coronal consonant context. This articulatory target – exhibited by both speakers presented here – is also accompanied by additional articulatory targets which vary by speaker: M.B. displays further constriction in the upper pharynx, while B.P. shows prepalatal constriction relative to coronal closure context. Despite differences in some aspects, these articulation patterns still result in acoustics which are more like /ɨ/ than lexical schwa, on average. However, the different certain aspects of articulatory targeting may help to explain differences in the observed patterns of acoustic variability for affix schwa. For instance, subject B.P. shows substantially less variability in F1 during affix schwa production and, at the same time, a prepalatal articulatory target which may lend stability to tongue height, which influences F1 (Fant, 1950a, b).

In addition to the lower pharyngeal target, it was also noted above that affix schwa can be characterized by a high tongue position relative to lexical schwa. Moreover, the analysis of targetedness presented here illustrates that, for subject B.P., this higher tongue position is actively controlled during that subject's production of affix schwa. No such target is observed, however, in subject M.B.'s data, despite the fact that that subject's tongue is especially high during affix schwa. This apparent discrepancy is potentially explained by remembering that the proposed methodology tests specifically for aspects of articulation that are actively controlled over and above any articulation that occurs in the consonantal and vocalic context. Thus, if M.B.'s tongue is already in a high position during the context segments, the test developed here will not reveal that position as targeted during the schwa. M.B.'s tongue may be high in the schwa context for several reasons. One possibility is that the tongue body, if it is otherwise unconstrained during the preceding coronal consonant, assumes a higher position in anticipation of the coming schwa (i.e., coarticulatory effects). Coarticulation may, indeed, be emphasized in this case due to the relatively shorter duration of M.B.'s affix schwa (fig. 2). On the other hand, it may be that M.B.'s production of /d/ already puts the tongue body into a high position. In this case, no further active control would be required during the schwa itself. It is not possible to differentiate between these two possibilities with the current data.

Despite this relatively clear evidence for articulatory targets in both types of schwa, as well as indications of what might be consistent about those targets across speakers, it is clear that more subjects are needed to overcome the large amount of interspeaker variability and to precisely characterize those targets. Wide interspeaker variability is observed for both lexical and affix schwa, both in the articulatory and acoustic domains. Collecting and analyzing data from a larger sample of subjects would clarify several questions. For example, it will be important to determine whether articulatory targets are unique to an individual or whether, perhaps, these 2 speakers are each representative of two classes of categorical targeted planning and behavior.

Attributing this interspeaker variability to specific factors, such as the physical structure (e.g., morphological) of the speech production apparatus or subtle dialectical variation, is also an important consideration.

It should be noted that the evidence for an articulatory target for affix schwa presented here is not entirely at odds with the hypothesis that the regular English past-tense suffix can be represented as a single coronal gesture whose coordination is determined by phonetic context (Goldstein, 2011). If this analysis of the past-tense suffix is correct, a vocoid could still result during the allomorph in question without the need to appeal to a distinct spatial target. The temporal coordination of consonant units would leave a temporal gap between the constrictions whose vocal tract shape is determined by the articulator motions resulting from the release of the initial constriction and by movement of individual articulators to the posture assumed during grammatical interspeech pauses. Those postures could, in fact, constitute a spatial target unto themselves. It has been observed that the speech articulators tend to return to particular postures during interspeech pauses. These postures are significantly different from, and less variable than, absolute rest postures and may therefore involve a higher degree of cognitive control (Ramanarayanan et al., 2013). Therefore, one possible way of accounting for the spatial target observed during affix schwa in the present study is that it represents the interspeech rest posture, rather than some phonologically meaningful target. It may be possible to examine this additional hypothesis empirically, but given the data collected for the present study, it must remain the domain of future work. Data from the present study is read speech and therefore consists almost entirely of continuous speech without interspeech pauses.

The methodology developed in the present study offers several advantages for addressing questions about articulatory targets. Performing analysis on a pixelwise basis avoids many of the common challenges related to processing high-dimensional image data, while allowing for very detailed examination of speech articulation kinematics. However, care needs to be taken in appropriate preprocessing of the image data and in formulating the analysis to ensure useful interpretation of the results. The test for targeted behavior presented here was carefully formulated to conform logically to a definition of targetedness based on predictability of behavior from context and coarticulation. Moreover, validation of this test on a small coronal consonant data set indicates that the test is sound. Validation of this test on larger and more diverse data sets will nonetheless be important for expanding its applicability to other studies. One key challenge in validating this test – or, indeed, any test of targeted behavior – is that there is no gold standard against which to compare the results. Articulatory targets of the kind investigated here have not been well established, which is also why the development of a computational test for targetedness is potentially useful. Still, reasonable validation of this or similar methods can be done, as in the present study, by comparing against expectations stemming from traditional descriptions of the phonetic segment in question.

Note that the methodology presented here, while providing a definition for targeted and targetless behavior, provides a test only for targeted behavior. The null hypothesis of our test for targeted behavior is, of course, that the behavior being tested is targetless. However, a nonsignificant result cannot necessarily be interpreted as evidence that the null hypothesis is true (i.e., that the behavior is targetless), but simply that there is little or no evidence that the behavior is targeted. Therefore, the correct

interpretation of those pixels not marked as targeted is that either they lack an articulatory target or alternatively that the effect of their articulatory target is too small for our test to confidently reject the null hypothesis.

5 Conclusion

Affix schwa, the vocoid associated with the English past-tense suffix, was examined in a variety of vowel contexts and compared to lexical schwa spoken in identical contexts. The examination was both articulatory and acoustic, including midsagittal vocal tract images from rtMRI with parallel, denoised audio. Speakers included 2 male American English speakers. The investigation included a characterization and comparison of formant frequencies, temporal duration, vocal tract posture and articulatory targetedness for both categories of schwa.

The use of rtMRI provides advantages in data quality stemming from complete midsagittal views of the vocal tract, including of the velum and pharynx. Moreover, image processing methods were described which are simple to implement and efficient to carry out, but which also take full advantage of this rich information by directly using pixel intensity variations as features describing speech articulation. A novel method for assessing targetedness was developed, including a way to test for statistical significance, as an extension of previous work to model motor action as a function of its surrounding context. These methods are widely applicable and may be of use in future studies of speech articulation and articulatory targets.

The present study provides evidence against the idea that affix schwa is targetless. Results do indicate, however, that affix schwa is different in many regards from lexical schwa, both acoustically and articulatorily. Affix schwa is generally higher than lexical schwa, both in vowel formant space and in tongue posture. However, the evidence presented here also indicates that the differentiating aspects of lexical and affix schwa vary across subjects, particularly in terms of articulatory targets. This is consistent with previous difficulties in characterizing affix schwa.

Acknowledgments

Work described in this article was supported by NIH Grant DC007124 as well as a graduate fellowship from the Annenberg Foundation.

Appendix A: Stimuli

Set	No.	Phrase	No.	Phrase
1	1	They rooted oozy plants	9	If Beta'd aped it once
	2	If Nota'd overheard	10	If Fonda'd offered more
	3	The fitted intervals	11	We bonded often then
	4	If Ruta'd oozed a lot	12	If Menda'd ever known
	5	If Fitta'd interviewed	13	If Needa'd even known
	6	If noted over lunch	14	He banded Annie's arm
	7	If mended ever more	15	The panda'd asked for more
	8	He baited apes a lot	16	If needed even once

Appendix A (continued)

Set	No.	Phrase	No.	Phrase
2	1	They rooted oozy plants	9	We bonded often then
	2	If Nota'd overheard	10	He baited apes a lot
	3	He banded Annie's arm	11	If Menda'd ever known
	4	If Needa'd even known	12	If Ruta'd oozed a lot
	5	If Beta'd aped it once	13	If noted over lunch
	6	If Fonda'd offered more	14	If mended ever more
	7	If Fitta'd interviewed	15	The fitted intervals
	8	The panda'd asked for more	16	If needed even once
3	1	If Fitta'd interviewed	9	We bonded often then
	2	If noted over lunch	10	They rooted oozy plants
	3	If Menda'd ever known	11	If Beta'd aped it once
	4	If mended ever more	12	If Nota'd overheard
	5	If Ruta'd oozed a lot	13	The panda'd asked for more
	6	If needed even once	14	If Needa'd even known
	7	He baited apes a lot	15	He banded Annie's arm
	8	If Fonda'd offered more	16	The fitted intervals
4	1	If noted over lunch	9	If Menda'd ever known
	2	The fitted intervals	10	If Ruta'd oozed a lot
	3	He baited apes a lot	11	If mended ever more
	4	He banded Annie's arm	12	They rooted oozy plants
	5	If needed even once	13	If Nota'd overheard
	6	If Needa'd even known	14	If Beta'd aped it once
	7	The panda'd asked for more	15	If Fitta'd interviewed
	8	We bonded often then	16	If Fonda'd offered more
5	1	If mended ever more	9	If Fitta'd interviewed
	2	They rooted oozy plants	10	If Ruta'd oozed a lot
	3	He banded Annie's arm	11	The fitted intervals
	4	He baited apes a lot	12	If Nota'd overheard
	5	If Beta'd aped it once	13	If Menda'd ever known
	6	If Fonda'd offered more	14	If noted over lunch
	7	We bonded often then	15	If needed even once
	8	The panda'd asked for more	16	If Needa'd even known
6	1	If Menda'd ever known	9	If Ruta'd oozed a lot
	2	If Beta'd aped it once	10	If Fitta'd interviewed
	3	He banded Annie's arm	11	If needed even once
	4	If Fonda'd offered more	12	If Needa'd ever known
	5	They rooted oozy plants	13	If mended ever more
	6	If noted over lunch	14	The panda'd asked for more
	7	If Nota'd overheard	15	The fitted intervals
	8	He baited apes a lot	16	We bonded often then
7	1	He baited apes a lot	9	The fitted intervals
	2	If Nota'd overheard	10	If noted over lunch
	3	If Fitta'd interviewed	11	The panda'd asked for more
	4	If needed even once	12	If Beta'd aped it once
	5	They rooted oozy plants	13	He banded Annie's arm
	6	If Ruta'd oozed a lot	14	If Fonda's offered more
	7	We bonded often then	15	If Menda'd ever known
	8	If Needa'd even known	16	If mended ever more

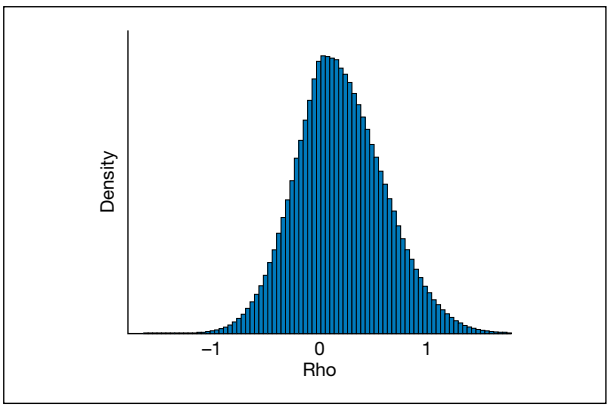

Fig. 6. Simulation results, showing the distribution of ρ resulting from |P| − |Q| taken over two million data points representing both P and Q. The standard deviations of P and Q were equal to 1.0 and 1.25, respectively. The variables P and Q were dependent, such that Pearson's r between their absolute values was equal to 0.83. The skewness of this distribution was equal to only 0.23.

Appendix B: Distribution of the ρ Statistic

In the present work, t tests are performed of the null hypothesis that the ρ statistic is distributed with mean zero, against the alternative hypothesis that the mean is greater than zero. It is believed that this choice of test is reasonable based on the following assumptions and reasoning. Suppose that X and Y are random variables representing the residual errors resulting from fitting the models 1 and 2, respectively, to articulatory data. Furthermore, the random variables X and Y have continuous distributions over R with probability density functions that are assumed to be normally distributed, approximately mean zero. Note that the random variables X and Y are likely dependent, since they are both constructed by subtracting the measured values of y. With the introduction of these new variables, the statistic ρ, presented in 3, can then be rewritten as $\rho = |X| - |Y|$. Applying the absolute value operation to X and Y (i.e., $|X|$ and $|Y|$) will transform the distribution of these variables, given the current assumptions, into half-normal distributions, which is a special case of the skew-normal distribution where the skewness parameter equals infinity. It has been shown that the negative of a skew-normal random variable is skew-normal, allowing for $-|Y|$, and that skew-normal random variables are closed under addition if they are dependent meaning that $|X| - |Y|$ remains skew-normal (Pouradmadi, 2007).

Therefore, there is reason to believe that ρ, though not normally distributed, belongs to a generalized class of distributions which includes the normal distribution. Moreover, it is expected that the skewness will be small because as the variance of X and Y becomes increasingly similar, the skewness of ρ will become very small. A simulation was conducted to illustrate this point. Two million data points were generated conforming to each of the two random variables, P and Q, with mean and skewness of 0, but P had a standard deviation equal to 1.0, while Q had a standard deviation equal to 1.25. The variables P and Q were dependent, such that the linear correlation (Pearson's r) between $|P|$ and $|Q|$ was equal to 0.83. The distribution of ρ resulting from $|P| - |Q|$ is shown in the histogram in figure 6. The skewness of this distribution was equal to only 0.23. The variances of residual errors produced by both models may, indeed, be very similar, as evidenced by their relatively similar fit overall (table 2), even if model 2 has an advantage. This, combined with a relatively large sample size (n > 50), lends confidence in using a t test in this situation.

References

Anderson S (1982): The analysis of French schwa. Language 58:535–573.
Bakovic E (2005): Antigemination, assimilation and the determination of identity. Phonology 22:279–315.
Bennett CM, Baird AA, Miller MB, Wolford GL (2010): Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for proper multiple comparisons correction. J Serendipitous Unexpected Results 1:1–5.
Boersma P (2001): Praat, a system for doing phonetics by computer. Glot Int 5:341–345.

Bresch E, Katsamanis A, Narayanan S (2010): Coupled HMM; in Proc Interspeech.
Bresch E, Narayanan S (2009): Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. IEEE Trans Med Imaging 28:323.
Bresch E, Nielsen J, Nayak K, Narayanan S (2006): Synchronized and noise-robust audio recordings during real-time MRI scans. J Acoust Soc Am 120:1791–1794.
Browman C, Goldstein L (1992): Targetless schwa: an articulatory analysis; in Docherty A, Ladd A (eds): Papers in Laboratory Phonology. Cambridge, Cambridge University Press, vol II: Gesture, Segment, Prosody, pp 26–56.
Browman C, Goldstein L (2000): Competing constraints on intergestural coordination and self-organization of phonological structures. Bull Commun Parlé 5:25–34.
Davidson L (2006): Phonotactics and articulatory coordination interact in phonology: evidence from nonnative production. Cogn Sci 30:837–862.
Devijver PA, Kittler J (1982): Pattern Recognition: A Statistical Approach. London, Prentice Hall.
Fant G (1950a): Transmission properties of the vocal tract. Part I. MIT Q Prog Rep, pp 20–23.
Fant G (1950b): Transmission properties of the vocal tract. Part II. MIT Q Prog Rep, pp 14–19.
Feise RJ (2002): Do multiple outcome measures require p-value adjustment? BMC Med Res Methodol 2:8.
Flemming E (2007): The phonetics of schwa vowels; in Minkova D (ed): Phonological Weakness in English. Palgrave.
Flemming E, Johnson S (2007): Rosa's roses: reduced vowels in American English. J Int Phonet Assoc 37:83–96.
Forsyth D, Ponce J (2002): Computer Vision: A Modern Approach. Englewood Cliffs, Prentice Hall.
Fromkin V (ed) (2000): Linguistics: An Introduction to Linguistic Theory. Oxford, Blackwell.
Gafos A (2002): A grammar of gestural coordination. Nat Lang Linguist Theory 20:269.
Geisser S (1993): Predictive Inference. New York, Chapman & Hall.
Goldstein L (2011): Back to the past tense in English; in Representing Language: Essays in Honor of Judith Aissen. Santa Cruz, Linguistics Research Center, University of California Santa Cruz.
Goldstein L, Byrd D, Saltzman E (2006): The role of vocal tract gestural action units in understanding the evolution of phonology; in Arbib M (ed): Action to Language via the Mirror Neuron System. Cambridge, Cambridge University Press.
Kiparsky P (1985): Some consequences of lexical phonology. Phonol Yearb 2:85–138.
Kitamura T, Takemoto H, Honda K, Shimada Y, Fujimoto I, Syakudo Y, Masaki S, Kuroda K, Oku-uchi N, Senda M (2005): Difference in vocal tract shape between upright and supine postures: observations by an open-type MRI scanner. Acoust Sci Technol 26.
Lammert A, Proctor M, Narayanan S (2010): Data-driven analysis of realtime vocal tract mri using correlated image regions; in Proc Interspeech.
Magen H (1989): An Acoustic Study of Vowel-to-Vowel Coarticulation in English; PhD thesis Yale University.
Nam H (2007): Articulatory modeling of consonant release gesture; in 16th International Congress of Phonetic Sciences.
Narayanan S, Nayak K, Lee S, Sethy A, Byrd D (2004): An approach to real-time magnetic resonance imaging for speech production. J Acoust Soc Am 115:1771–1776.
Perneger TV (1998): What's wrong with Bonferroni adjustments. BMJ 316:1236–1238.
Pinker S, Prince A (1988): On language and connectionism: analysis of a parallel distributed processing model of language acquisition; in Pinker S, Mehler J (eds): Connections and Symbols. Cambridge, MIT Press, pp 73–193.
Pouradmadi M (2007): Construction of skew-normal random variables: are they linear combinations of normal and half-normal? J Stat Theory Application 3:314–328.
Ramanarayanan V, Goldstein L, Byrd D, Narayanan S (2013): An investigation of articulatory setting using real-time magnetic resonance imaging. J Acoust Soc Am 134:510–519.
Rothman KJ (1990): No adjustments are needed for multiple comparisons. Epidemiology 1:43–46.
Smorodinsky I (2002): Schwas with and without Active Gestural Control; PhD thesis Yale University.
Stone M, Stock G, Bunin K, Kumar K, Epstein M, Kambhamettu C, Li M, Parthasarathy V, Prince J (2007): Comparison of speech production in upright and supine position. J Acoust Soc Am 122:532–541.
Tiede M, Masaki S, Vatikiotis-Bateson E (2000): Contrasts in speech articulation observed in sitting and supine conditions; in Proc International Seminar on Speech Production, Bavaria, pp 25–28.
Trager A, Smith A (1951): An Outline of English Structure. Studies in Linguistics Occasional Papers (No. 3). Norman, Battenberg Press.
Wrench A, Cleland J, Scobbie J (2011): An ultrasound protocol for comparing tongue contours: upright vs. supine; in Proceedings of the International Congress of Phonetic Sciences, Hong Kong, pp 2161–2164.
Zsiga E (2003): Articulatory timing in a second language. Stud Second Lang Acquisition 25:399–432.