Truncation of Pharyngeal Gesture in English Diphthong [aɪ]

Fang-Ying Hsieh[1], *Louis Goldstein*[1], *Dani Byrd*[1], *Shrikanth Narayanan*[2]

[1] Department of Linguistics, University of Southern California, USA
[2] Viterbi School of Engineering, University of Southern California, USA

fangyinh@usc.edu

Abstract

It is well acknowledged that [a] in English diphthongs (e.g. [a] in "pie'd") has a different formant structure from its closest corresponding monophthong (e.g. [a] in "pod"). The current study proposes that these two sounds share the same cognitive unit, i.e. the pharyngeal constriction gesture that produces [a], and the surface difference can be modeled as a consequence of truncating the same articulatory movement in time by the following palatal glide in the diphthongal environment. Formation of pharyngeal constriction gesture during the production of [a] in a diphthong and in its corresponding monophthong was observed in various timing contexts using Realtime MRI; and the collected production data were quantitatively analyzed using the direct image analysis (DIA) technique, which infers tissue movement by tracking pixel intensity change over time in regions of interest. Results support our truncation account in that: (1) formation time of pharyngeal constriction is significantly longer in monophthongs than in diphthongs; (2) this duration correlates with the resulting constriction degree; and (3) the resulting constriction degree predicts the acoustic difference in the F2 dimension as predicted by our hypothesis.

Index Terms: English diphthongs, speech production, rtMRI

1. Introduction

Diphthongs are typically viewed as formant movement from one vowel target to another in the same syllable [1, 2]. Nevertheless, acoustic studies have shown that the formants of the initial and terminal vowels of a diphthong are not necessarily the same as the simple vowels (which occur as monophthongs) used to describe them. Instances can be found in English: the initial vowel in diphthong [aj] (as in "pie") has been customarily transcribed with a central low vowel [a], while its closest corresponding monophthong (as in "pod") has been transcribed with back low vowel [ɑ]. Nevertheless, positing distinct phonological units for these two elements based on the acoustic difference ignores the fact that these differences are not arbitrary and could potentially be accounted for within a model of speech production. In particular, it is possible to hypothesize that a single cognitive unit of articulatory control is engaged during the production of both low vowels, but the differences emerge due to the spatio-temporal context of other units in which it is embedded.

In line with the above rationale, we propose that these two [a]'s share the same phonological unit, i.e. the pharyngeal constriction gesture that produces [a], and that the surface acoustic difference results from truncation of the gesture in time in the production of [a] by the following palatal glide in diphthongs. The proposal is couched in the theoretical framework of Articulatory Phonology [3], in which articulatory gestures are the basic units of contrast among lexical items as well as units of articulatory action. Gestures are defined spatially and temporally by sets of parameters that specify the task (with respect to constriction location and constriction degree) and time course of the gesture's unfolding. Unlike other phonological theories that depend on abstract symbols (phonemes) as the mental representation of contrastive lexical units, in Articulatory Phonology, the representation using articulatory gestures predicts the actual physical processes and the resulting sounds that constitute speech.

In order to test our hypothesis about the acoustic difference in [a]'s being a lawful consequence of the temporal interval available to form the target pharyngeal constriction, production data were collected using the real time magnetic resonance imaging (rtMRI) technique. This allows us to directly observe the tongue body movement in the pharyngeal area, which is not possible using other techniques like electromagnetic articulography (EMA) [4] or X-ray microbeam [5]. Our hypothesis predicts that the formation time of pharyngeal constriction should be longer in monophthongs than in diphthongs, due to onset of the production of palatal constriction for the [j]; and this duration will be a predicator of the resulting pharyngeal constriction degree of [a], which in turn predicts the surface acoustic difference. To be more specific, we predict that the longer formation time of pharyngeal gesture in monophthong will result in more constricted pharynx and lower F2 (which is associated with vowel backness), compared with that in diphthong [aj].

2. Method

2.1. Data acquisition

Midsagittal speech articulation from two adult female speakers of American English (Speaker A and Speaker B) was captured using a custom MR protocol [6, 7] with denoised audio [8]. Both speakers read the designed sentences eight times. Two repetitions from Speaker B were not included in data analysis due to a different pulse sequence being used when collecting those two repetitions, yielding different spatial (and temporal) resolutions from the rest. As our analysis technique relies on selecting a specific pixel intensity region for comparison across tokens, these two repetitions cannot be combined with the rest. Data were reconstructed into a sequence of video frames with spatial resolution of 68-by-68 pixels, with pixels 3 mm in width, at a temporal rate of 23.18 frames per second.

Two factors influencing timing were manipulated for the target monophthongs and diphthongs: 1) sentence position: final (pre-IP boundary) versus non-final (pre-ip or smaller boundary) and 2) syllable coda: none, voiced coda or voiceless. These factors have been reported [9, 10] to affect vowel length: sentence-final position and a voiced coda

lengthen the vowel, while the non-final position and a voiceless coda shorten the vowel. The designed carrier sentences are illustrated below.

Sentence-final position: I didn't think I'd see _____.

Non-final position: I didn't think I'd see _____ anymore.

The carrier phrases were designed with a tongue-fronting gesture (palatal constriction /i/ in "see") preceding the target pharyngeal constriction gesture (which retracts tongue body) in order to clearly reveal the onset of the time course of pharyngeal constriction formation.

All target words are monosyllabic and begin with a labial (/p/ or /f/) or an alveolar (/t/) gesture, ending in an alveolar (/t/ or /d/) or labial (/p/ or /m/) gesture. Diphthongs were produced in the condition without a following coda, but monophthong /a/ was not. Two fillers "tot" and "Todd" were added to balance the number of sentences in each block and also for the other purpose of future study. Only the labial-initial words have been analyzed in the current study. The complete list of target words utilized for the MRI scan is illustrated in the table below:

Table 1. *List of target words*

Coda type	Labial-initial		Alveolar-initial	
	Monoph-thong	Diph-thong	Monoph-thong	Diph-thong
Voiceless	*pot*	*fight*	*top*	*type*
Voiced	*pod*	*pied*	*Tom*	*time*
No coda	--	*pie*	--	*tie*
Fillers	--		*tot, Todd*	

All the 12 target words in Table 1 occur in both sentence-final and non-final positions, yielding 24 sentences in total. Sentences are divided into four blocks, each containing six sentences. The subjects repeated the whole set of sentences in fixed order eight times during the scan.

2.2. Data analysis

In order to compare the pharyngeal constriction degree in monophthongs and diphthongs, the collected MR images were automatically analyzed using the direct image analysis technique, henceforth DIA [11, 12, 13]. DIA uses the mean pixel intensities within regions-of-interest to infer local articulator movement based on the fact that regions of high intensity correspond to the presence of soft tissue or tissue compression in that region, while regions of low intensity correspond to areas of air. As the tongue body retracts in forming a pharyngeal constriction, higher intensity in the pharyngeal region would imply the presence of more tongue tissue and thus more constricted pharynx. Applied to other regions of the vocal tract, this method has been shown to successfully capture the constriction location and constriction degree of Italian singleton and geminate consonants [14]. The alternative to DIA would be to determine the vocal tract constriction degree through automatic segmentation of images along the air-tissues boundaries [15]. However, such a method tends to require intensive computation and produces noisy outcomes that need manual correction. In the current study, a semi-automatic segmentation method [16] was employed on a subset of the data in order to find the optimal intensity analysis region in the pharynx and to test the validity of using DIA.

2.2.1. Selecting the optimal pharyngeal region

The optimal intensity analysis region was chosen by finding the region within which the intensity values most accurately reflect individual token differences in pharyngeal constriction degree measured as the 2D area of air space in the pharynx resulting from (manually corrected) segmentation analysis. A quarter of data for each subject was selected to apply both the segmentation analysis and DIA.

The segmentation method uses a speaker-specific semi-polar composite analysis grid superimposed on each image to be segmented and identifies the tissue boundaries by seeking pixel intensity thresholds along the gridlines. The complete grid extends from the glottis to the region anterior to the lips. The air space enclosed between the two gridlines adjacent to epiglottis and uvula was of our interest. An example of the superimposed analysis grid and the region of interest (enclosed by the two yellow lines) for one subject can be seen in Figure 1(a).

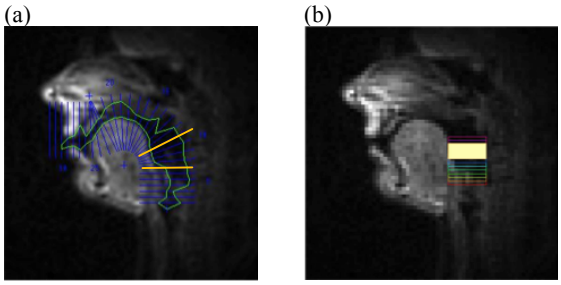

Figure 1: *Illustration of image analysis for Speaker A. Images show (a) the superimposed analysis grid and region of interest for segmentation analysis; (b) candidate analysis regions for DIA analysis and the final analysis region being selected.*

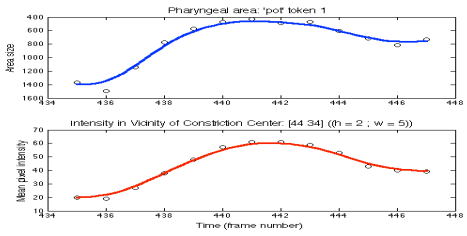

Figure 2: *Examples of time functions of constriction degree (the first repetition of "pot" in final position produced by Speaker A) resulting from the two analysis methods.* Top*: time function of pharyngeal air space resulting from segmentation analysis;* Bottom: *time function of mean intensity change in the region shown in Figure 1(b).*

As previously mentioned, the segmentation results could be noisy and thus were manually corrected for each image frame. The size of this air space was measured in pixels and used as the indicator of constriction degree: the smaller the air space, the more constricted the pharynx is. This area change was tracked during the production of target words, yielding a time function for each token (e.g. the top graph in Figure 2).

For finding the optimal intensity analysis region in the pharynx used for DIA, we first picked a set of pixels parallel

to the pharyngeal wall. With each of these pixels being the center, a rectangle area was delimited and served as one analysis region candidate. The width of the delimited region was chosen such that it covers the maximal air space in the pharynx when the tongue body was pulled forward, and thus the region always included some portion of tongue root so as to track tongue tissue movement at all time. The candidate regions used for Speaker A are illustrated by the unfilled rectangles with different edge colors in Figure 1(b). This is a temporal frame captured during the production of /i/ in the phrase "see fight." As the tongue body has not yet retracted, each analysis region has low intensity at this time point. Averaging pixel intensity of a region for the image sequences during the production of a pharyngeal constriction yields an intensity time function, exemplified in Figure 2(b). The collected intensity functions and the area time functions were both smoothed using a locally weighted linear regression [17] in order to eliminate the random fluctuations across frames.

The pixel region for which the intensities across tokens correlate most highly with pharyngeal areas at maximal constriction (independently defined as intensity maximum and pharyngeal area minimum for each token) was chosen as the optimal pixel analysis region (e.g. the filled rectangle in Figure 1b was chosen for Speaker A. At the time points of maximal constriction, the correlation between the pixel intensities and the area sizes is strong ($r < -0.78$) across selected tokens for both subjects. Moreover, within each token, the time functions of pixel intensities and the pharyngeal areas also have strong correlation ($r < -0.75$) for both subjects. These results demonstrate the validity of using DIA to depict the pharyngeal constriction process.

2.2.2. Measurements

After separately selecting the analysis region for each subject based on the correlations between two methods applied onto the selected quarter of data, DIA on the selected regions was extended to all the data. As intensities of these regions were shown to be capable of capturing pharyngeal constriction degree, the resulting constriction degree for each pharyngeal gesture was measured as the peak intensity in the time function.

To test our hypothesis about truncation of constriction being the result of reduction in time available to form the pharyngeal constriction, durations were measured from the time of release of the preceding palatal constriction gesture (i.e. the gesture that produces /i/ in "see") to the achievement of pharyngeal constriction (defined as the first temporal point in the smoothed function after it exceeds a designated 90% threshold of within-token range). The palatal constriction was also defined using DIA. But for choosing analysis region, we simply adopted the method utilized in [14] rather than the approach we used for capturing pharyngeal constriction: The site with maximal dynamic range of intensity for most tokens was used for tracking palatal constriction over time. The release of palatal constriction was defined as the first temporal point in the smoothed function after the smoothed palatal intensity time function falls lower than the designated threshold (90 %) of the maximum intensity.

Acoustic analysis was also conducted for this study: average F1 and F2 during the interval of pharyngeal constriction (defined as the time during which the pharyngeal intensity time function is above 90% of the intensity range) were measured using Praat [18].

3. Results

Analysis results show that monophthongs and diphthongs are significantly different ($p < 0.05$ for two-sample t-test) in formation duration (i.e. the interval from the release of preceding palatal constriction to the achievement of pharyngeal constriction), constriction degree, and in F2 for both subjects. There is also a significant difference between monophthongs and diphthongs in F1 for Speaker B, but not for Speaker A. These results are shown in Table 2 below; and F1 and F2 values of each token are plotted in Figure 3.

Table 2. *Measurement results. (Dur = formation duration; Const = constriction degree defined by localized mean pixel intensities).*

Means	Speaker A			Speaker B		
	Mono	Diph	Sig	Mono	Diph	Sig
Dur (ms)	176	159	*	206	180	*
Const	55	50	*	49	45	*
F1 (Hz)	994	974	--	1143	1088	*
F2 (Hz)	1439	1680	*	1710	1880	*

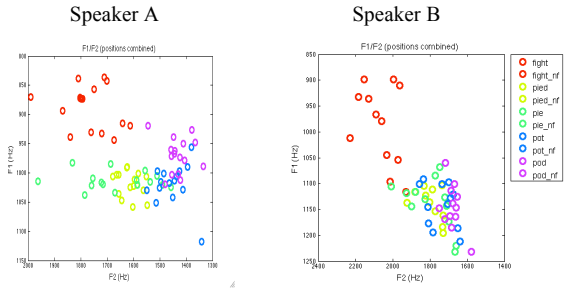

Figure 3: *Acoustic analysis results: vowel formants for collected data*

Since our proposal predicts a dependency among the formation time, resulting constriction degree, and acoustics, pair-wise correlation coefficients were computed for these variables. Significant correlations ($p < 0.05$) were found among these variables, as shown in Table 3. Since monophthongs and diphthongs are consistently different in F2 across speakers but not in F1, F1 was not included.

Table 3. *Correlation coefficients*

R	Speaker A	Speaker B
Duration vs. CD	0.75*	0.67*
CD vs. F2	-0.68*	-0.71*
F2 vs. Duration	-0.34*	-0.48*

We can also compare the pharyngeal constriction formation trajectory over time in monophthongs and diphthongs by tracking the change in pixel intensity in the pharyngeal region during the formation time. Averaging token intensities for each target word in either final or non-final position over time results in the trajectory contours shown in Figure 4.

Speaker A Speaker B

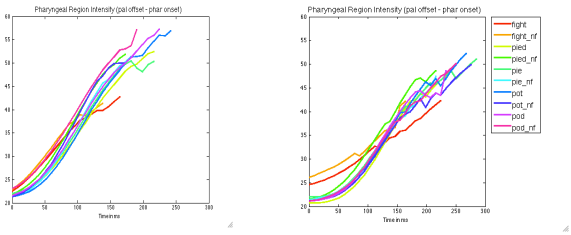

Figure 4: *Averaged pharyngeal intensity change during the constriction formation interval for each target word. (Note that the ends of some lines are noisy due to unequal length of individual tokens.)*

4. Discussion

Results in section 3 support our hypothesis that the acoustic difference between [a] in monophthongs and diphthongs could be a consequence of the time interval available in forming pharyngeal constriction: significant correlations were found between formation time, resulting constriction degree, and F2 (Table 3). The correlations are strong in general, but not for the relationship between formation duration and F2. This slightly weaker correlation is not unexpected, however, because the predicted causal relationship is such that the (articulatorily defined) duration of movement controls constriction degree, and the constriction degree determines the resulting F2**.** The indirect causal relation between duration and F2 predicts that the formation time will only explain a percentage of variance in F2 that is accounted for by constriction degree. Thus the resulting lower correlation between duration and F2 is reasonable.

The acoustic analysis results (Figure 3) revealed speaker-specific differences in the acoustics of produced [a]: Speaker A produced [a]'s that vary mainly along the F2 dimension, but not along the F1 dimension (except for the tokens of target word "fight," which will be discussed later); while the [a]'s produced by Speaker B vary in both dimensions. These results may suggest a need to scrutinize the anterior portion of vocal tract in addition to fully understand the relationship between production and acoustics of low vowel [a].

On the other hand, examining the constriction formation trajectory based on the regional pixel intensity change over time (in Figure 4) shows that in general the formation of pharyngeal constriction that produces [a] progresses similarly in monophthongs and diphthongs, as shown by the similar slopes. These similar trajectories support our proposal that the [a] in diphthongs and monopthongs share the same cognitive unit—the same pharyngeal constriction gesture, but this gesture is truncated in time in diphthongs, resulting in reduced constriction degree, as indicated by the correlation results.

However, an exception is found: the pharyngeal constriction gesture in diphthong [aj] when followed by a voiceless consonant (i.e. "fight" in final and non-final positions) has a different constriction formation trajectory slope, and has a higher pharyngeal constriction degree than the rest at the very beginning of the formation interval. This difference is small for Speaker A but apparent for Speaker B. In addition, vowels of this word on the formant charts (Figure 3) show that it is distinctive in F1. This exceptional behavior may be explained by the Canadian Raising sound change found in some dialects of the northern United States, in which case the initial vowel in diphthongs [aj] and [aw] is produced as a higher vowel before voiceless consonants than in other environments [19]. This variant is often transcribed as a central mid vowel [ʌ] in the literature. Accordingly, a distinctive gesture may be posited when accounting for this variant.

It is interesting to note, however, that the slopes of pharyngeal intensity trajectories forming [a] in final and non-final "fight" are only slightly different from the others for Speaker A, while their formation duration is consistently shorter than the others; on the contrary, for Speaker B, the slopes are intrinsically different from the rest, but not the duration. This might suggest that a distinctive gesture is necessarily to be posited for [a] in "fight" for Speaker B, but not for Speaker A. Besides the radically different slope, a similar formation interval results in different constriction degree in the case of Speaker B. However, for Speaker A, the slightly different slope and distinctive duration might still suggest the same pharyngeal gesture, since the distinctive acoustics could be a consequence of the distinctively short formation time.

Leaving out tokens of target word "fight" in final and non-final positions for both speakers, the correlations between formation duration and resulting constriction degree, and that between constriction degree and F2 remain significant, (though they are weaker) but not for the correlation between F2 and constriction duration. As previously discussed, this could be due to the indirect relationship between those two measures.

5. Conclusions

Overall, the similar constriction formation trajectories and significant correlation between articulation and acoustics support our hypothesis that the acoustically different [a] in monophthongs and diphthongs share the same phonological unit, i.e. the pharyngeal constriction, and suggest that the acoustic difference can be modeled as a consequence of truncation in time. However, the current proposal has not addressed the cause of truncation in diphthongs. In particular, it remains unexplained why a coda consonant (e.g. /d/ in "pod") does not truncate the pharyngeal constriction as a palatal glide (e.g. /j/ in "pie") does. One possibility is the nucleus-coda difference (between /j/ and /d/), another is the strong bio-mechanical coupling between the pharyngeal and glide gesture (both control the tongue body) that necessitates the early turn-off of the pharyngeal gesture. We expect to answer this question by exploring the timing of production of consonant codas and glides in our collected data in the near future.

In addition, the strong correlation between the pharyngeal air space area and intensity time function for both subjects demonstrates that DIA is capable of quantifying the differences in constriction degree when applied to pharyngeal constriction gestures. The automatic and comparatively efficient computation of this technique would be beneficial in further studies of vocal tract dynamics.

6. Acknowledgements

Work described in this paper was supported by NIH Grants DC007124, DC008780.

7. References

[1] Holbrook, A. and Fairbanks, G., "Diphthong formants and their movements," J. Speech Hear. Res. 5, 33-58, 1962.

[2] Lehiste, I., and Peterson, G. E. "Transitions, glides and diphthongs," JASA, 38: 268-277, 1961.

[3] Browman, C. P., & Goldstein, L., "Dynamics and articulatory phonology," In *Mind as motion: Dynamics, behavior, and cognition*, R. Port, & T. van Gelder, Eds., Cambridge, MA: MIT Press, 1995.

[4] Perkell, J.S., M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," JASA, 92(6): 3078–3096, 1992.

[5] Westbury, J., G. Turner, and J. Dembowski, *X-ray microbeam speech production database user's handbook*. Waisman Center on Mental Retardation and Human Development, University of Wisconsin, 1994.

[6] Narayanan, S., Nayak, K., Lee, S., Sethy, A. and Byrd, D., "An approach to real-time magnetic resonance imaging for speech production", JASA, 109: 2446, 2004.

[7] Narayanan, S., Bresch, E., Ghosh, P., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A., Proctor, M. , Ramanarayanan, V. and Zhu, Y., "A Multimodal Real-Time MRI Articulatory Corpus for Speech Research", in Proc. Interspeech, 2011.

[8] Bresch, E., Nielsen, J., Nayak, K. and Narayanan, S., "Synchronized and noise-robust audio recordings during realtime MRI scans", JASA, 120: 1791, 2006.

[9] Jones, D. *An outline of English phonetics* (9th ed.). Cambridge: W. Heffer & Sons Ltd, 1972.

[10] Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J., "Segmental Duration in the Vincinity of Prosodic Phrase Boundaries," JASA 91(3): 1707-1717, 1992.

[11] Bresch, E., Katsamanis, A., Goldstein, L. and Narayanan, S., "Statistical multi-stream modeling of real-time MRI articulatory speech data," in INTERSPEECH: 1584–1587, 2010.

[12] Lammert, A., Proctor, M. and Narayanan, S., "Data-Driven Anal- ysis of Realtime Vocal Tract MRI using Correlated Image Re- gions", in INTERSPEECH: 1572–1575, 2010.

[13] Proctor, M., Bone, D., Katsamanis, N., & Narayanan, S., "Rapid Semi-automatic Segmentation of Real-time Magnetic Resonance Images for Parametric Vocal Tract Analysis," Proc. Interspeech, Makuhari Messe, Japan, 2010.

[14] Hagedorn, C., Proctor, M., Goldstein, L. & Narayanan, S., "Automatic Analysis of Geminate Consonant Articulation using Real-time MRI," ISSP'11, Montreal QC, 20-23 June 2011.

[15] Bresch, E. and Narayanan, S., "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images", IEEE Trans. Med. Imaging, 28(3): 323, 2009.

[16] Proctor, M., Lammert, A., Katsamanis, A., Goldstein, L., Hagedorn, C., and Narayanan, S., "Direct Estimation of Articulatory Kinematics from Real-time Magnetic Resonance Image Sequences", In INTERSPEECH: 281-284, 2011.

[17] Atkeson, C., A. Moore, and S. Schaal**,** "Locally weighted learning," AI Review, 11: 11–73, 1997.

[18] Boersma, P., "Praat, a system for doing phonetics by computer," *Glot International* 5:9/10, 341-345, 2001.

[19] Britain, D., "Dialect contact and phonological reallocation: "Canadian raising" in the English Fens," Language in Society 26 (1): 15–46, 1997.