Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Speech Communication
Session 5aSCb: Production and Perception II: The Speech Segment (Poster Session)

5aSCb27. Pharyngeal constriction in English diphthong production

Fang-Ying Hsieh*, Louis Goldstein, Dani Byrd and Shrikanth Narayanan

***Corresponding author's address: Linguistics, University of Southern California, Los Angeles, CA 90089, fangyinh@usc.edu**

This study tests the hypothesis that the acoustic difference between [a] in English diphthongs (e.g. [a] in "pie'd") and its corresponding monophthong (e.g. [a] in "pod") results from the same pharyngeal gesture being truncated by the following palatal glide in the diphthongal environment. Production data were collected with real-time MRI and have been analyzed using the direct image analysis (DIA) technique, which infers tissue movement by tracking pixel intensity change over time in regions of interest. Preliminary results show that (1) DIA is capable of capturing the timing and magnitude of the pharyngeal constriction gesture which produces [a]; and (2) the proposed hypothesis is supported: the formation time of pharyngeal constriction strongly correlates with the resulting constriction degree as predicted.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4799762]
Received 22 Jan 2013; published 2 Jun 2013

INTRODUCTION

In early acoustic studies, diphthongs were described as formant movement from one vowel target to another (Lehiste and Peterson 1961, Holbrook and Fairbanks 1962). Nevertheless, it is also generally acknowledged that the formants of the initial and terminal targets are not necessarily compatible with the closest corresponding monophthongs used to describe them. For instance, the initial vowel of /aɪ/ can vary from /a/ to /æ/; and the final vowel can be /ɪ/ to /ɛ/. While the terminal vowel can be considered as a syllable coda, and its variation can be attributed to the more general coda-weakening phenomenon in American English (Gick 2002), the varying nature of the initial vowel in diphthongs remains unexplained. The current study seeks to account for the acoustic difference between the initial vowel in a diphthong and its corresponding monophthong as a function of the temporal coordination of compositional units of speech production.

We propose a hypothesis to account for this difference: while the initial vowel has the same constriction target as its closest corresponding monophthong, this acoustic difference in the initial vowel results from truncation of the vowel gesture by the following distinct glide gesture. To test this hypothesis, we observe the pharyngeal constriction gesture during the production of [a] in a diphthong and in its corresponding monophthong in different contexts that influence timing, using the real time magnetic resonance imaging (rtMRI) technique. Our hypothesis predicts that if the acoustic difference is a consequence of truncating the same pharyngeal constriction gesture, the resulting pharyngeal constriction degree of [a] should be predicted by the temporal interval available to form the constriction, and the differences in constriction degree should in turn predict the acoustic differences.

EXPERIMENTAL METHOD

Data Collection

An adult female English-Spanish bilingual speaker was imaged while producing designed sentences containing diphthongs or monophthongs using a custom MR Imaging protocol (Narayanan et al. 2004). The subject's vocal tract was imaged midsagittally, with a spatial resolution of 68 x 68 pixels (200 x 200 mm.) and a temporal reconstruction rate of 33.18 frames per second.

Two factors influencing timing were manipulated for the target monophthongs and diphthongs: 1) sentence position: final versus non-final and 2) syllable coda: none, voiced coda or voiceless. These factors have been shown to affect vowel length: sentence-final position and a voiced coda lengthens the vowel, while the non-final position and a voiceless coda shortens 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.

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). The target diphthongs include /aɪ/ and /aʊ/; only /aɪ/ diphthongs have been analyzed in the current study. The stimuli used in this study are listed in Table 1.

TABLE 1. Stimuli for the experiment

Coda type	**Labial-initial**		**Alveolar-initial**	
	Monophthong	**Diphthong**	**Monophthong**	**Diphthong**
Voiced Coda	pod	pie'd	top	type
Voiceless Coda	pot	fight	Tom	time
No Coda	--	pie	--	tie

Each of the above stimuli occurs in both final and non-final sentential contexts. The subject repeated the whole set of sentences in fixed order twice during the scan.

Data Analysis

In order to estimate the constriction degree of the pharynx, we employ an automatic method—direct image analysis, henceforth DIA (Lammert et al., 2010). DIA infers articulator movement directly based on pixel intensity 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. This method has been shown to successfully capture the constriction location and constriction degree of Italian singleton and geminate consonants (Hagedorn et al. 2011). As tongue body retracts in forming a pharyngeal constriction, higher intensity in the pharyngeal region implies the presence of more tongue tissue and thus more constricted pharynx.

Another method to determine constriction degree is by segmenting the images along the air-tissue boundaries. While this method can also be automatic (Bresch and Narayanan 2009), it requires intensive computation and produces noisy outcome that needs manual correction. For the current study, the segmentation method is used supplementarily to find the optimal intensity analysis region in the pharynx and to ensure the validity of using DIA.

To begin with, a set of pixels parallel to the pharyngeal wall were selected. With each of these pixels as the center, a horizontal rectangle area is delimited as a candidate analysis region. The delimited region was chosen to always include some portion of tongue root so as to fully track tongue tissue movement. An example is given in Figure 1 below. The average pixel intensity of this region is taken as an indicator of the degree of pharyngeal constriction. This is a temporal frame during the production of /f/ in the phrase "see fight." As the tongue body has not retracted, the analysis region has low intensity at this time point.

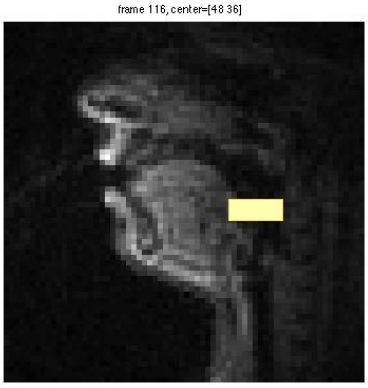
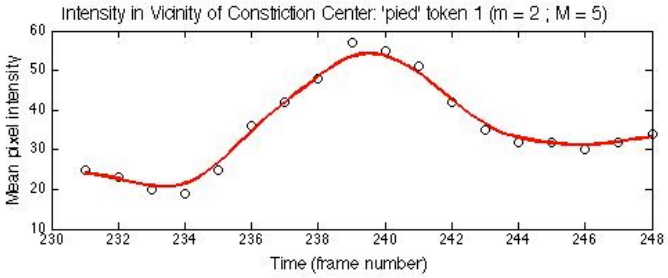


FIGURE 1. Intensity analysis region of pharyngeal constriction. This is a temporal frame during the production of /f/ in the phrase "see fight."

Averaging pixel intensity of the analysis region for the image sequences during the production of target words (c.f. Table 1) yields intensity time functions, exemplified in Figure 2a. The intensity functions (and the area time function exemplified in Figure 2b) used in this study were smoothed using a locally-weighted linear regression (LWR) (Atkeson et al., 1997) in order to eliminate the random fluctuations across frames. The data were upsampled by a factor of 5 (between frames) and the smoothing window width was 0.9 frame (27.09 ms.).

(a)



(b)

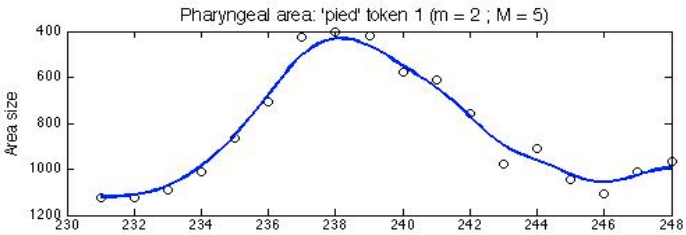


FIGURE 2. Example of resulting pharyngeal intensity time function (a) and area time function (b): the first token of "pie'd" in sentence-final position. The peaks in both figures represent maximal constriction.

For the same image sequences, we also delineated the contour of the air space in pharynx using a semi-automatic segmentation method developed in Proctor et al. (2010). This method uses a 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 uvular was of our interest. As previously mentioned, the segmentation results could be noisy and thus were manually corrected. The size of this air space was measured in pixels and used as another indicator of constriction degree: the smaller the air space, the more constricted the pharynx is. This area change was also traced over time, yielding another time function for each token (e.g. Figure 2b). The area time functions served as a reference for selecting the optimal pixel region for intensity analysis: the pixel region for which the intensities across token correlate most highly with pharyngeal areas at maximal constriction was chosen as the optimal pixel analysis region.

RESULTS

The rectangular region illustrated in Figure 1 above was chosen as the optimal analysis region for this speaker based on the criterion mentioned in last section. At the time points of maximal constriction (independently defined by the two methods as maximal intensity and minimal pharyngeal area) the correlation between the pixel intensities and the area sizes is strong ($r = -0.69$) across all tokens. Moreover, within each token, the time functions of pixel intensities and the pharyngeal areas also have strong correlation ($r < -0.85$). These results ensure the validity of using DIA to depict pharyngeal constriction process.

Our hypothesis is that the truncation of constriction is the result of reduction in time available to form the pharyngeal constriction. Thus formation time should be a predictor of constriction degree across the various timing environments. We define formation time as the time between the achievement of target in the onset gesture (labial or alveolar) to the time of achievement of the pharyngeal constriction. The latter is defined as the first temporal point in the smoothed function after the intensity time function exceeds the designated threshold (90 %) of the maximum intensity. A threshold is given instead of using the time point at which the peak intensity occurs because many tokens (especially those containing a monophthong) show a plateau in the time function.

For the gestures involved in onset consonants, as Hagedorn et al. (2011) has shown DIA is capable of capturing constriction location and constriction degree in labial and alveolar consonants, we adopted the same method to find the constriction location and constriction degree of the onset consonant. The time point at peak intensity is designated as the achievement of target of the onset consonant, thus the onset time of the formation of the pharyngeal constriction.

With formation time thus defined, correlation analyses were conducted between the formation duration and the resulting constriction degree (defined as peak intensity values). The labial-initial and alveolar-initial tokens are analyzed separately in that the formation duration was measured with a different starting point (lip gesture vs. tongue tip gesture). As predicted, the correlation between the formation time and resulting constriction degree is strong for both sets of data, including both monophthongs and diphthongs. The resulting coefficients are listed in Table 2.

TABLE 2. Coefficients of the correlation between formation time and constriction degree

Data type	Coefficients
Labial-initial tokens	0.69
Alveolar-initial tokens	0.61

CONCLUSIONS

The strong correlation between formation time and resulting constriction degree of pharyngeal constriction gesture across monophthongs and diphthongs in different timing environments supports our hypothesis and suggests that the acoustic difference in [a] between monophthongs and diphthongs could be a consequence of truncation in time. However, other possible hypotheses still need to be tested to clarify the cause of this truncation. For instance, why a consonantal coda (e.g. /d/ in "pod") does not truncate the pharyngeal vowel [a] as a glide does (e.g. /j/ in "pie"). One possibility is due to the nucleus-coda difference (between /j/ and /d/), or due to different tract variables involved. To further solidify our proposal, we anticipate collecting more data and conducting additional analyses. Additional analyses include systematic analysis of the formants, comparing the trajectory of constriction formation in monophthongs and diphthong, and a detailed comparison among the timing environments.

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

ACKNOWLEDGMENTS

Research supported by NIH Grant.

REFERENCES

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

Bresch E. and S. Narayanan **(2009).** "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Trans. Med. Imag*. 28(3), 323-338.

Gick, B. **(2002).** "An X-ray investigation of pharyngeal constriction in American English schwa," Phonetica, 59, 38–48.

Hagedorn, C., Proctor, M., Goldstein, L. & Narayanan, S. **(2012).** "Automatic Analysis of Geminate Consonant Articulation using Real-time MRI." ISSP'11, Montreal QC, 20-23 June 2011.

Holbrook, A. and Fairbanks, G. **(1962).** "Diphthong formants and their movements," J. Speech Hear. Res. 5, 33-58.

Lammert, A., M. Proctor and S. Narayanan **(2010).** "Data-driven analysis of realtime vocal tract mri using correlated image regions." In *Proceedings of Interspeech 2010*. Mahuhari, Japan.

Lehiste, I., and Peterson, G. E. **(1961).** "Transitions, glides and diphthongs," J. Acoust. Soc. Am. 38, 268-277.

Narayanan, S., K. Nayak, S. Lee, A. Sethy, and D. Byrd. **(2004)**. "An approach to real-time magnetic resonance imaging for speech production," J. Acoust. Soc. Am , 115(4), 1771–1776.

Proctor, M., Bone, D., Katsamanis, N., & Narayanan, S. **(2010)**. "Rapid Semi-automatic Segmentation of Real-time Magnetic Resonance Images for Parametric Vocal Tract Analysis." Proc. Interspeech, Makuhari Messe, Japan, 26-30 Sept. 2010: 1576-1579.