EVIDENCE FOR PIVOTS IN TONGUE MOVEMENT FOR DIPHTHONGS

Boram Kim[1,2], Mark K. Tiede[2], D. H. Whalen[1,2,3]
[1]City University of New York Graduate Center; [2]Haskins Laboratories; [3]Yale University

ABSTRACT

Tongue movement from one speech element to another has been found to consist largely of two patterns, "pivot" and "arch" [5]. Diphthongs act as a single element for some purposes but two for others. What does the pattern of tongue changes indicate about their status? Our data come from ultrasound, which lacks information on most of the hard structures. This limitation requires a new method to quantify patterns of tongue movement. Tongue shapes of diphthongs and the vowel sequences with syllable (Korean) or mora (Japanese) boundaries were collected with the HOCUS system [14] from native speakers of Mandarin, Korean and Japanese. The sequence of tongue gestures from each token was contoured. Computational modelling of the strength of the pivoting movement indicate that diphthongs consist of two elements, with pivot and arch patterns, similar to previous results, and possibly a new "shift" pattern, in which the entire visible surface moves.

Keywords: Pivots, Diphthongs, Methodological research, Speech Production

1. INTRODUCTION

Although tongue movement in speech is habitually controlled by speakers, its kinematics are complex. Iskarous [5] proposed that tongue movement and the change of area function from one speech element to the other show two main systematic kinematic patterns, termed the "pivot" and the "arch". The pivoting pattern is found when the tongue gesture of two segments make maximal movements in two constriction locations and minimal movement between them. At the pivot point, very little change occurs in midsagittal distance function (i.e., the distance from tongue to palate and posterior pharyngeal wall at vocal tract). The arching pattern is characterized by having a single region where the midsagittal distance function changes, and another region has very little change. The pivot pattern has been found to be relevant to perception [6, 11]. Additional studies have found pivoting in Russian [13] and in clinical populations [8]. Although these studies supported the findings from [5], studies on dynamic patterns of tongue movement are very limited. More importantly, the methodological issue that an X-ray database of speech analysed in [5] estimated midsagittal distance function which is a crucial component of area function [3, 15] based on a polar-rectangular grid. This method requires an extensive view of both the soft and the hard structures (i.e., the edge of the tongue, palate and posterior pharyngeal wall). Because the calculation of pivots requires a quantification of the majority of the tongue surface, point parameterizations such as x-ray microbeam [7] or electromagnetic articulometry [12] do not allow assessment of pivots. Ultrasound provides a relatively extensive view of the tongue shape, but the images lack the referential hard structures. This limitation of ultrasound images, which are widely used in articulatory study, requires new methods to quantify tongue movement if we are to use the data for assessing the pivot hypothesis.

The current paper proposes a method of quantifying the degree of pivoting in tongue kinematics based on tongue gestures from ultrasound images. Tongue kinematic patterns from diphthongs and the corresponding sequence of vowels produced by native speakers of three different languages (i.e., Mandarin, Korean and Japanese) were investigated. Acoustic studies suggest that diphthongs have two elements [4, 9]. The second formant, in particular, plays an important role as its rate of transition varies among diphthongs and influence their identification [2]. Whether those two elements are related by a pivot pattern is undetermined; Iskarous [6] called [ai] a "sequence," so diphthongs may differ.

The present study uses ultrasound images to a) investigate whether the previously proposed two basic patterns of tongue movement could explain the majority of the tongue movement found in diphthongs and vowel sequences from three languages; b) whether the pivoting patterns could be quantified based on only tongue gestures, and c) whether diphthongs might exemplify a new pattern of tongue movement which does not fall into two basic patterns.

2. EXPERIMENT

2.1 Participants

6 native speakers, 1 male and 1 female speaker each, of Seoul Korean, Tokyo Japanese and Standard (Mandarin) Chinese participated the experiment. None reported having been diagnosed with any language or speech disorder.

2.2 Stimuli

For each language, the target words were written in their own writing system. Participants were instructed to produce 5 repetitions of 3 target words. The target words for Mandarin speakers consisted of three Mandarin diphthongs. To elicit the diphthong /ei/, the target word /bei/ was presented to participants, but they were asked to pronounce only the vowel of the target word and omit the onset consonant. The target words for Korean and Japanese consisted of sequences of two vowels. The Korean vowel sequences all contained syllable boundaries (marked by dots) between two vowels resulting in two syllables for each vowel sequence. The Japanese vowel sequences all contained mora boundaries which are marked by hyphens. Although there is a mora boundary, the Japanese vowel sequences are considered to be a single syllable. Table 1 shows all target words from each language.

Table 1: Stimuli list.

Languages	Targets	Characters	Meaning
Mandarin	/ai/	哀	'sadness'
	/au/	凹	'concave'
	/(b)ei/	杯	'cup'
Korean	/a.i/	아이	'child'
	/a.u/	아우	'younger brother'
	/e.i/	에이	'non-word'
Japanese	/a-i/	愛 あい	'love'
	/a-ɯ/	会う あう	'to meet'
	/e-i/	エイ えい	'fish'

2.3 Ultrasound recording

An ultrasound system (Ultrasonix; Sonix Touch) using a basic 2D imaging mode was used to collect real-time mid-sagittal images of the tongue. The center frequency of the ultrasound transducer was set to 6.5 MHz. The field of view was 148° (sector 75% = 111°), and an imaging depth was 8 cm. The frame rate was 59.9 frames per second, which provides a sufficient quantity of time-frames for the analysis of the diphthongs. The microconvex ultrasound transducer was placed underneath the participant's chin with a fixed spring-loaded transducer holder. The transducer holder was fixed to a weighted customized pedestal.

2.4 Ultrasound image contouring

The start and the end of the target vowels were demarcated from the audio recording based on the waveform and the spectrogram using Praat software [1]. Every frame from each vowel interval was contoured using EdgeTrack [10], which computes 100 x-y coordinates and these coordinates were then resampled to be equally-spaced and ordered such that the first point is the most anterior tongue point. The Haskins optically-corrected ultrasound system (HOCUS) was used to correct for probe movement relative to the head by remapping contours to a head-centric coordinate system [14].

2.4 Quantifying the patterns of tongue movement

An algorithm was devised to quantify pivot strength. Tongue contours were analyzed by a short time window with a fixed window length *w*, sliding through all time frames (time step = 1 frame), resulting in *N* analysis windows. A window contains tongue contours in 2**w*+1 time frames (*w* preceding and *w* following time frames, centered at the temporal (frame) index *i*). Fig. 1 demonstrates an example of tongue movements of the Mandarin diphthong /ai/, starting from the red lines and ending in the blue lines. Fig. 2a indicate the tongue contours across 2**w*+1 frames, centered at the *i* frame, extracted from the total frames in Fig. 1.

For each window, an anchor point (the black circle in Fig. 2a and 2b) is defined as the point on the tongue with the smallest average distance to all points from the other frames, taking into account the 2**w*+1 frames in the current window. The local maximal tongue displacements immediately more anterior and posterior to the anchor point are defined as the pre-context (*C1*) and the post-context (*C2*), respectively (Fig. 2a-b, where the green circle and line indicate *C1* and the brown ones indicate *C2*). A pivot strength S_i, for the i^{th} window, is defined as the mean of $C1_i$ and $C2_i$. The pivot frame is defined as the frame containing the pivot with the largest S_i (in this example, between the blue and red lines of Fig. 2a).

The tongue movement is considered as having a pivot pattern when both *C1* and *C2* are above a threshold *k* (default = 10% maximum tongue displacement), and they are both relatively large, i.e., their absolute difference ||*C1* - *C2*|| is below *k* (Fig. 2). An example of the pivot pattern analysis shown in Figs. 2 is a token of the Mandarin diphthong /ai/ with *w* = 4 (window length = 150ms). As shown in Fig. 2b, the tongue displacements from *C1*, passing the anchor, to *C2* formed a V-shape pattern for the pivot tongue movements.

Figure 1: Example of pivot pattern in the Mandarin diphthong /ai/. Thick red line indicates the start of the diphthong; thick blue line the end of the diphthong.

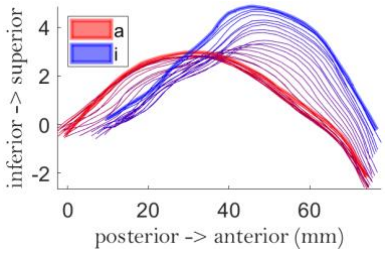
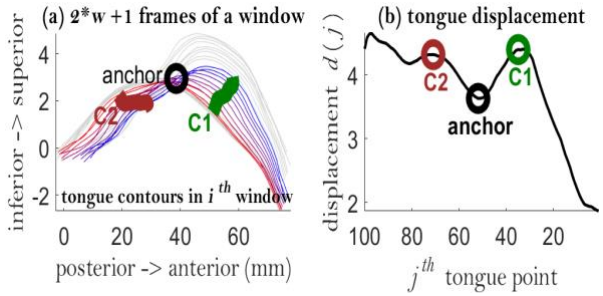


Figure 2: Tongue contours across frames of an analysis window (a) and the corresponding tongue displacement for each tongue point (b).



An example of the arching pattern from the Korean vowel sequence /a.u/ is shown in Figs. 3 and 4, with w = 6 (window length = 217ms). An arching pattern requires that either *C1* or *C2* is above *k* and their difference is relatively large, i.e., ||*C1* - *C2*|| is also above *k* (Fig. 4b). In Fig. 4b, the tongue displacement plot for arching pattern shows that only *C1* is above the threshold *k*. The tongue displacements from *C1* to *C2* shaped a diagonal pattern for the arch tongue movements.

As physiological differences (e.g., vocal tract size) across speakers might influence maximum tongue displacement, pivot strength is normalized within each speaker using a z-transformation. Also, each speaker differs in speech rate, which could influence the number of frames per vowel. Thus, the window length is presented as the portion of the total number of frames (NFrame) for each target vowel (2**w*+1 /NFrame * 100).

Figure 3: Example of arching pattern in the Korean vowel sequence /a.u/. Thick red line indicates the start of the vowel; thick blue line the end of the vowel.

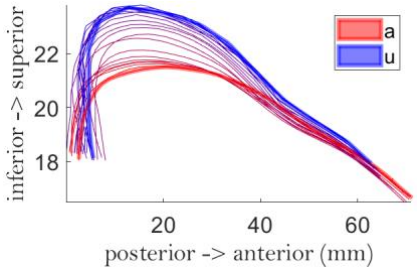
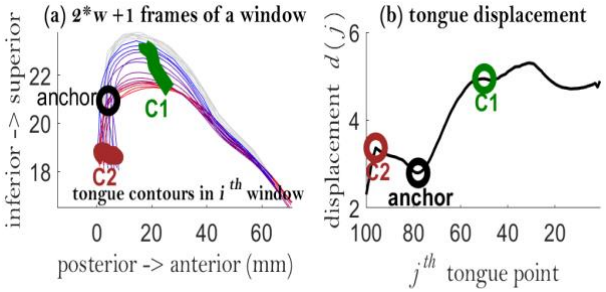


Figure 4: Tongue contours across frames of an analysis window (a) and the corresponding tongue displacement for each tongue point (b).



3. RESULTS

3.1. Pivot strength

Table 2 presents the optimal window sizes (*w*) and maximum pivot strengths (*S*) for the diphthong/vowel sequences /ai/, /ei/ and /au/, averaged across the three languages. Among the three vowels, /ai/ has the strongest pivot strength (3.1) followed by /au/ and /ei/. Table 3 shows the optimal window sizes and pivot strengths for each language, averaged across vowels. The Korean vowel sequence have similar pivot strength and smaller optimal window size than Mandarin diphthongs. This suggests that although Korean and Mandarin diphthong/vowel sequences have similar pivoting tongue displacements, the pivoting tongue movements in the Mandarin diphthong extends longer in time than those in Korean. On the other hand, Japanese seems to have weaker pivot strength showing lesser tongue displacements in pivoting tongue movements. Table 4 shows that the values do not differ much by for each language. Lastly, Table 5 summarizes the pivot strengths and optimal window size for each vowel in each language.

Table 2: Optimal window size and normalized maximum pivot strength by vowels.

	/ai/	/ei/	/au/
window size (%)	76	49	54
pivot strength	3.1	1.5	2.2

Table 3: Optimal window size (*w*) and normalized maximum pivot strength (*S*) by language.

	Diphthongs	vowel sequence	
	Mandarin	Korean	Japanese
w (%)	68	51	89
S	2.9	2.9	1.7

Table 4: Optimal window size (*w*) and normalized maximum pivot strength (*S*) by speaker.

	Mandarin		Korean		Japanese	
	F	M	F	M	F	M
w (%)	37	68	68	51	73	27
S	2.7	2.9	2.6	2.9	2.1	2.1

Table 5: Optimal window size (*w*) and normalized maximum pivot strength (*S*) of vowels by language.

vowels		Mandarin	Korean	Japanese
/ai/	*w*	68	76	95
	S	2.9	3.1	2.2
/ei/	*w*	54	22	65
	S	2.5	2.2	1.4
/au/	*w*	56	21	54
	S	0.3	2.1	2.2

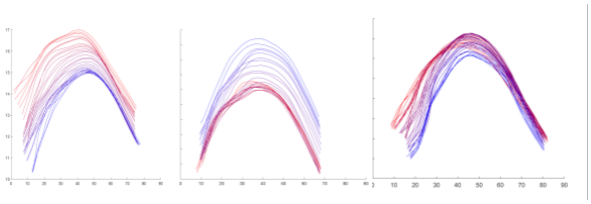
Besides pivoting tongue movements, most of the arching movements were found in the Mandarin diphthong /ei/ and the Korean vowel sequence /a.u/.

3.2 Shifting pattern

There are, nonetheless, tongue movements that cannot be identified either as pivot or arch patterns by our algorithm. Most of those unidentifiable tongue movements show a consistent shifting pattern (Fig. 5). The shifting patterns were found in around 20% of our dataset; they were all from either /ei/ or /au/ tokens (except for one token from /ai/). Unlike the pivot or arching patterns, the shifting pattern shows more vertical tongue movement. Shifting patterns were found in all speakers, but there were individual differences. For example, the Mandarin female speaker used mostly shifting patterns for /au/ but arching patterns for /ei/, whereas for the Mandarin male speaker, the shifting patterns were primarily used for both /au/ and /ei/. Similarly, the Japanese female speaker used the shifting patterns for some of /a-ɯ/ and /e-i/ tokens, while the Japanese male speaker mostly used pivoting pattern for /a-ɯ/. Korean speakers showed a shifting pattern in the /e.i/ sequence.

However, one limitation of using ultrasound to image tongue gestures is that the tongue tip and tongue root are not visible in ultrasound image. It is possible that there is a pivoting or arching point at the tongue tip or tongue root. The frequency and, indeed, existence of this shift pattern needs further investigation.

Figure 5: shifting patterns of Mandarin, Japanese and Korean /ei/ vowels.



4. DISCUSSION AND CONCLUSION

For three languages (Mandarin, Japanese and Korean), diphthongs and vowel sequences were found to have many examples of pivot and arch patterns [5]. A method of calculating pivot strength in the absence of quantification of the hard structures of the vocal tract was found to be useful. Our data indicate that diphthongs may make common use of a third category of tongue kinematics, the "shift" pattern. It remains to be seen whether this pattern is due primarily to tongue movement per se or rather to changes in jaw height.

The relatively complex kinematics of tongue movement appear to be the result of a limit on the number of dynamic patterns actively used in speech, as proposed by Iskarous [5]. The current results suggest that there may be a third pattern—shift, as well as pivot and arch—that is commonly used. Quantifying the strength of the pivots from easily obtained ultrasound images holds the promise of more direct exploration of individual differences (typical and atypical) and language-specific differences in articulatory dynamics.

5. ACKNOWLEDGEMENTS

This research was supported by NIH DC-002717 grant to Haskins Laboratories. We thank Wei-rong Chen for comments that greatly improved the manuscript. We thank two anonymous reviewers for their insights.

6. REFERENCES

[1] Boersma, P. and Weenink, D., *Praat: Doing phonetics by computer [Computer software]. Version 6.0. 33*. 2017.

[2] Gottfried, M., Miller, J. D., and Meyer, D. J., Approches to the classification of American English diphthongs, *Journal of phonetics*, vol. 21, no. 3, pp. 205–229, 1993.

[3] Heinz, J. M. and Stevens, K. N., On the derivation of area functions and acoustic spectra from

cineradiographic films of speech, *The Journal of the Acoustical Society of America*, vol. 36, no. 5, pp. 1037–1038, 1964.
[4] Holbrook, A. and Fairbanks, G., Diphthong formants and their movements., *Journal of Speech and Hearing Research*, vol. 5, p. 38, 1962.
[5] Iskarous, K., Patterns of tongue movement, *Journal of Phonetics*, vol. 33, no. 4, pp. 363–381, 2005.
[6] Iskarous, K., Nam, H., and Whalen, D. H., Perception of articulatory dynamics from acoustic signatures, *The Journal of the Acoustical Society of America*, vol. 127, no. 6, pp. 3717–3728, 2010.
[7] Kiritani, S., X-ray microbeam method for measurement of articulatory dynamics-techniques and results, *Speech Communication*, vol. 5, no. 2, pp. 119–140, 1986.
[8] Kocjančič, T., Ultrasound and acoustic analysis of lingual movement in teenagers with childhood apraxia of speech, control adults and typically developing children, PhD Thesis, Queen Margaret University, 2010.
[9] Lehiste, I. and Peterson, G. E., Transitions, glides, and diphthongs, *The journal of The acoustical society of America*, vol. 33, no. 3, pp. 268–277, 1961.
[10] Li, M., Kambhamettu, C., and Stone, M., Automatic contour tracking in ultrasound images, *Clinical linguistics & phonetics*, vol. 19, no. 6–7, pp. 545–554, 2005.
[11] Nam, H., Mooshammer, C., Iskarous, K., and Whalen, D. H., Hearing tongue loops: Perceptual sensitivity to acoustic signatures of articulatory dynamics, *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3808–3817, 2013.
[12] Perkell, J. S., Cohen, M. H., Svirsky, M. A., Matthies, M. L., Garabieta, I., and Jackson, M. T., Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements, *The Journal of the Acoustical Society of America*, vol. 92, no. 6, pp. 3078–3096, 1992.
[13] Proctor, M., Towards a gestural characterization of liquids: Evidence from Spanish and Russian, *Laboratory Phonology*, vol. 2, no. 2, pp. 451–485, 2011.
[14] Whalen, D. H., Iskarous, K., Tiede, M. K., and Ostry, D. J., HOCUS: The Haskins optically-corrected ultrasound system for measuring speech articulation, *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2632–2632, 2004.
[15] Yehia, H., Rubin, P., and Vatikiotis-Bateson, E., Quantitative association of orofacial and vocal-tract shapes, in *Audio-Visual Speech Processing: Computational & Cognitive Science Approaches*, 1997.