ULTRASOUND STUDY OF GESTURAL TIMING IN MANDARIN VOWEL-NASAL PRODUCTION

Ya Li & Sonya Bird

University of Victoria, British Columbia, Canada
yali@uvic.ca; sbird@uvic.ca

ABSTRACT

The present study uses ultrasound in combination with acoustic analysis to examine how a tongue gesture is timed in relation to acoustic syllable duration in Mandarin vowel-nasal production. The results show that gestural timing is relevant to both tongue and syllable positions; for example, anterior gestures such as the tongue tip rising tend to occur closer to syllable peripheries than open ones such as the tongue dorsum lowering, and toward which periphery they occur depends on which syllable position they hold. While this timing pattern can be explained by a biomechanically based model of gestural organization, others appear to be perceptually linked.

Keywords: ultrasound, gestural timing, Mandarin vowel-nasal production

1. INTRODUCTION

This study seeks to understand the relationship between articulatory organization and syllable structure. Previous studies find that characteristic patterns of gestural organization are associated with syllable position. For example, English /l/ is found to have a more backed articulation in syllable final than initial position, and in syllable final position, the tongue tip gesture of /l/ occurs later than its tongue dorsum gesture [3, 4]. Although these findings provide an articulatory basis by which to define syllables, it is still questionable as to why these patterns occur and whether or not they are language specific [3].

The Jaw Cycle Hypothesis (JCH) proposed by Redford [6] offers a biomechanically based answer to the first question by claiming that the regular open-close motion of the jaw provides a structural frame for gestures to fall into place according to their degree of jaw opening: gestures associated with a closed jaw occur during the closed portion of a jaw cycle while those with an open jaw occur during the open portion of a jaw cycle. Here the closed and open portions of a jaw cycle are respectively associated with syllable peripheral (i.e., onset or offset) and central (i.e., nuclear) positions. The correspondence between tongue position and jaw openness is illustrated in Figure 1:

Figure 1: Correspondence between tongue position and jaw openness (adapted from Redford [6]).

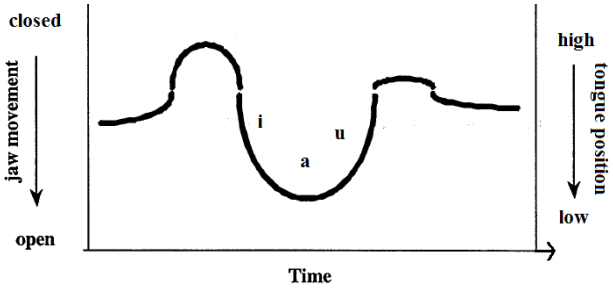


Figure 1 shows that, as the jaw moves up and down, the tongue changes its position from high to low as a function of time. Different segments are plotted vertically against their relative jaw height and tongue position and horizontally according to their syllable position. For example, /a/ (including [a, ɑ]) is produced with an open jaw and also the tongue dorsum lowering or retraction, so it is located at the bottom of the contour. /i/ before and above /a, u/ indicates that /i/ is located at syllable onset and the tongue body rising gesture of /i/ is associated with a more closed jaw and/or higher tongue position than the tongue dorsum gestures of /a, u/. Similarly, /u/ after and above /a/ indicates that /u/ is located at syllable offset and associated with a more closed jaw and higher tongue position than /a/.

If jaw motion is a phonetic universal, then the associated gestural patterns should not be language specific. In order to find some evidence for this assumption, the present study uses Mandarin data to examine the relative timing of 3 types of tongue gestures, anterior, raised, and open, in the production of 4 types of Mandarin syllables, V, GV, VN, and GVN. Here *V*, *G*, and *N* respectively refer to vowel, glide, and nasal; *Anterior*, *raised*, and *open* respectively refer to the tongue tip/body rising, the tongue dorsum rising, and the tongue dorsum lowering/retraction. Note that the term

raised, adopted from Esling [2], is similar to but more descriptive than the traditional term *high, back*. Based on the JCH, this study predicts that anterior gestures occur closer to syllable peripheries than raised/open ones.

2. METHODOLOGY

2.1. Participants and speech material

Four female native Mandarin speakers participated in this study. The speech material is a word list containing 24 attested monosyllabic Mandarin words, and they correspond to 5 V, 5 GV, 6 VN, and 8 GVN syllables. The Mandarin *Pinyin* and phonetic representations (in square brackets) of these words are listed in Table 1:

Table 1: Twenty-four Mandarin (G)V(N) syllables.

V	*yi* [i]	*yu* [y]	*wu* [u]	*a* [a]	*e* [ɤ]
GV	*ya* [ja]	*yue* [ɥɛ]	*wo* [wo]	*wa* [wɑ]	*ye* [jɛ]
VN	*in*[in]	*ün*[yn]		*an*[an]	*en*[ən]
	ing[iŋ]			*ang*[ɑŋ]	
GVN	*yan* [jɛn]	*yuan* [ɥɛn]		*wan* [wan]	*wen* [wɤn]
	yang [jɑŋ]		*yong* [juŋ]	*wang* [wɑŋ]	*weng* [wɤŋ]

Note that the high level tone was chosen for all the test words because its long duration may be able to help sustain a tongue gesture for easy observation in the ultrasound video.

2.2. Data collection

This study used a GE Logiq-e portable ultrasound machine with a 3.5MHz electronic convex intercoastal transducer, 8C-RS, to collect tongue image data. Speakers sat comfortably in a high chair with their heads leaning slightly forward to allow their chins to rest on the transducer. The transducer was fixed to a microphone stand.

The test words were randomly presented on a computer screen in front of the speakers. Each test word gradually appeared 4 times (hence 4 tokens for each word) in a PowerPoint slide with a 2-second interval between every two appearances. Speakers were asked to read each token as it appeared on the screen. This on-screen, time-controlled word-reading method, since its first adoption in Li's [2] acoustic study of vowel-nasal production, has proved to be an effective way to control the reading speed and hence minimize coarticulatory variations caused by different speech rates.

The tongue image data were collected by the ultrasound machine and transmitted into an EMS RGB XTreme PCI-e video capture card equipped computer. The speech sounds were simultaneously recorded with a microphone and transmitted into the same computer via an M-audio Pre-Amp. UltraCap, a custom video/audio capture program developed by Coey [1], was used to control the video/audio data recording process. Note that the video/audio streams, especially the end portions of the two streams, sometimes were not aligned (i.e., synchronized) due to technical limitations, so this study took necessary measures such as reversing word order and repeating recording sessions to make sure that no video/audio misalignment is perceivable in the final data. As a result, the alignment issue shows little effect on timing patterns, as they are fairly consistent across different tokens and speakers.

2.3. Data analysis

This study used the following equation (1), the percentage of Ar_D over Ac_D, represented by Ar_T%, to indicate gestural timing, that is, how soon a gesture occurs in relation to acoustic syllable duration.

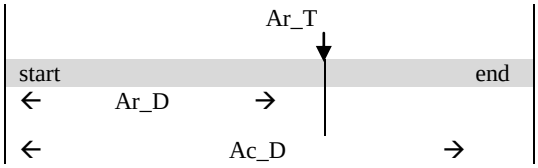
$$(1) \quad \text{Ar\_T\%} = \frac{\text{Ar\_D}}{\text{Ac\_D}} \times 100\%$$

Ar_T is the time when a gesture reaches its peak (i.e., the maximum excursion of the tongue) during the syllable production. For example, Ar_T would be measured at the point of the maximum tongue body rising for /i/ and at the point of the maximum tongue dorsum lowering or retraction for /a/ ([a] or [ɑ]). Figure 2 illustrates 3 video frames containing 3 tongue positions during the production of /i/: the tongue body starting to rise (left), rising to the highest position or gesture peak (center), and starting to pull back (right). The 3 frames are 0.33s apart. Note that the tongue surface contour is highlighted from the tongue root to tongue body.

Figure 2: Three tongue positions of /i/.

Ar_D is the duration from the acoustic start of the syllable to Ar_T, and Ac_D is the acoustic duration of the syllable. Figure 3 illustrates the timing relationship among Ar_T, Ar_D and Ac_D.

Figure 3: An illustration of Ar_T, Ar_D, and Ac_D.



If Ar_T% is 10% and 50% respectively for an anterior and an open gesture, then the anterior gesture is said to occur earlier than the open one, whether they are in the same syllable such as in /ja/ or in different syllables of the same type such as in /in, an/ (the VN type). In other words, for a given syllable position, the gesture timing of a segment should depend mainly on tongue position, regardless of its syllable affiliation.

The tongue imaging data were visually examined frame by frame in Sony Vegas Pro 8.0. Once a gesture peak was identified in a video frame, the time corresponding to the frame was taken as Ar_T. Note that not all tokens were consistently produced or clearly imaged. To make sure that each word has an equal number of usuable tokens, only 2 out of 4 tokens (mostly the 2nd and 3rd token) were analyzed for each word.

The audio data were manually segmented in Praat 5.1.31. The start and end points of each syllable were identified based on their waveforms and spectrograms. Note that the use of acoustic rather than articulatory syllable boundaries to infer gestural timing requires synchronization between video and audio streams but avoids the difficulty segmenting syllables in the continuous ultrasound video stream.

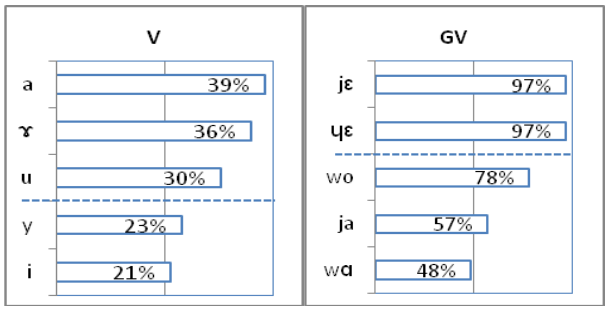
Then Ar_D, Ac_D, and Ar_T% were calculated for each gesture in each chosen token and averaged across the 2 chosen tokens and 4 speakers. Finally, the averaged Ar_T% was compared for different gestures in different syllables of the same type. The results of the comparison are used to inform gestural timing patterns. Due to the small amount of data, no statistical analysis is performed on the results, so the timing patterns reported here are only preliminary.

3. RESULTS AND DISCUSSION

3.1. Gestural timing for vowels

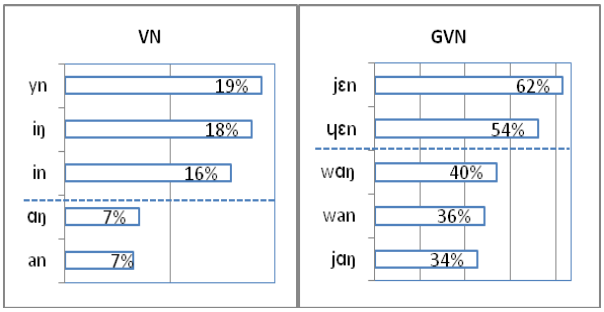
The Ar_T% results for vowels suggest that vowel quality and syllable type play a major role in determining when tongue gestures for vowels occur over acoustic syllable duration. Chart 1 below shows that anterior gestures occur more toward syllable onset in V production and more toward syllable offset in GV production than raised/open ones, as evidenced respectively by the smaller Ar_T% for /i, y/ than for /u, ɤ, a/ (see bars below and above the dotted lines on the left) and the larger Ar_T% for [ɥɛ, jɛ] than for [wo, ja, wɑ] (see bars above and below the dotted lines on the right).

Chart 1: Comparisons of the Ar_T% for V in (G)V production.



The gestural timing patterns in (G)V production support the prediction of this study: anterior gestures occur more syllable peripherally than raised/open ones. Chart 2, however, illustrates some timing patterns not predicted by this study.

Chart 2: Comparisons of the Ar_T% for V in (G)VN production.



In Chart 2 (left), open gestures occur very early, even earlier than anterior ones in VN production, as evidenced by the smaller Ar_T% for [an, ɑŋ] than for [in, iŋ, yn] (see bars below and above the dotted line). In Chart 2 (right), anterior gestures occur later than open ones in GVN production, as evidenced by the larger Ar_T% for [ɥɛn, jɛn] than for [jaŋ, wan, wɑŋ] (see bars above and below the dotted line). These timing patterns appear to go against the prediction: open gestures should occur more syllable centrally than anterior ones.

The early occurence of open gestures in VN production suggests that the acoustic start of /a/ is roughly aligned with the maximum jaw opening. Such alignment makes /a/ perceptually prominent at the very beginning, rendering the syllable /an/ as [aãn] in the actual production. If the jaw opening was not wide enough at the acoustic start, the syllable /an/ could be heard as [əãn]. In fact, perceptually motivated gestural timing is found to be more often assoicated with syllable initial position than the JCH-based timing [3].

Note that anterior gestures do not occur as early as open ones in VN production perhaps because whether the tongue body rises early (presumably for vowels /i, y/) or very early (presumably for glides /j, ɥ/) will not affect the perception of /in, yn/, as long as the jaw remains closed. Speakers may have opted for the production of /in/ as [iĩn] rather than [jĩn] simply for the ease of articulation.

The occurrence of anterior V gestures toward syllable central position in GVN production may well be another case with perceptual motivation. Unlike other GVN syllables, [ɥɛn, jɛn] contain only anterior gestures. Because the jaw is hinged at the back and opens at the front, the quality of front vowels is a function of jaw position, not of lingual setting [2]. As a result, the perception of [jɛn], for example, may depend mainly on jaw opening.

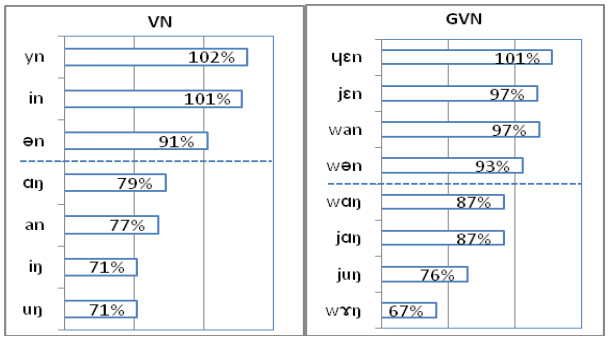
If the jaw opened too soon from [j] to [ɛ], then the transition from [j] to [ɛ] would be short, rendering the second part of [jɛn] more salient than the first part. Then the syllable [jɛn] would be heard as a heavily nasalized sequence [jɛ̃n] rather than the actual sequence [jɛɛ̃n]. Therefore, the perceptual salience of [j] and [ɛ] may have to be achieved by distancing the gesture peak of [ɛ] from that of [j].

3.2. Gestural timing for nasals

The Ar_T% results for nasals suggest that nasal place plays a major role in determining when oral closure gestures for nasals occur over acoustic syllable duration. Chart 3 below shows that the oral closure gesture tends to occur later for front /n/ than for back /ŋ/ in both VN (left) and GVN (right) productions, as evidenced by the smaller Ar_T% for /ŋ/ than for /n/ in most cases (see bars below and above dotted lines).

Given that the oral closure gesture of /n/ is an anterior gesture involving the tongue tip rising and the oral closure gesture of /ŋ/ is a raised gesture involving the tongue dorsum rising, the gestural pattern for nasals also supports the prediction of this study: anterior gestures occur more toward syllable offset than raised ones.

Chart 3: Comparisons of the Ar_T% for N in (G)VN production.



4. CONCLUSION

By examining gestural timing in Mandarin (G)V(N) production, this study suggests that anterior gestures tend to occur more peripherally than raised/open ones. This pattern supports the JCH-based prediction in that tongue (hence jaw) position plays a major role in gestural timing. However, the early occurrence of open gestures in VN production and the late occurrence of anterior V gestures in GVN production suggest that perceptual factors may override the effect of tongue position on gestural timing.

5. REFERENCES

[1] Coey, C. 2009. *UltraCap (Version 1.4).* [Computer program]. University of Victoria.

[2] Esling, J. 2005. There are no back vowels: The laryngeal articulator model. *Canadian Journal of Linguistics* 50(1), 13-44.

[3] Gick, B., Campbell, F., Oh, S., Tamburri-Watt, L. 2006. Toward universals in the gestural organization of syllables: A cross-linguistic study of liquids. *Journal of Phonetics* 34(1), 49-72.

[4] Krakow, R.A. 1999. Physiological organization of syllables: A review. *Journal of Phonetics* 27, 23-54.

[5] Li, Y. 2008. *Mandarin Speakers' Production of English and Mandarin Post-Vocalic Nasals.* MA Thesis, University of Victoria.

[6] Redford, M.A. 1999. *An Articulatory Basis for the Syllable.* Ph.D. dissertation, University of Texas, Austin.