On the inter-dependence of tonal and vocalic production goals in Chinese

Jason A. Shaw[1,4], Wei-rong Chen[2], Michael I. Proctor[3], Donald Derrick[1], Elita Dakhoul[1]

[1]*MARCS Institute, University of Western Sydney, Australia*
[2]*National Tsing Hua University, Taiwan*
[3]*Macquarie University, Australia*
[4]*School of Humanities and Communication Arts, University of Western Sydney, Australia*

J.Shaw@uws.edu.au g944710@oz.nthu.edu.tw michael.proctor@mq.edu.au D.Derrick@uws.edu.au

Abstract

We studied tone-vowel coproduction using Electromagnetic Articulography (EMA). Fleshpoints on the tongue and jaw were tracked while native Chinese speakers (n = 6) produced three vowels, /a/, /i/, /u/, combined with four Chinese tones. We found differences in tongue position across tones for /a/ and for /i/ but not for /u/. The low and rising tones patterned together in conditioning lower tongue blade (TB) position for /a/ and a higher TB position for /i/. This pattern suggests a degree of inter-dependence between tonal and vocalic targets. The effect of tone on TB height was mediated by jaw movement such that, even as TB sensor position varied across tones, the Euclidean distance between TB and Jaw sensors within each vowel remained stable. Thus, for this set of Chinese vowels, there is a relational invariance between active articulators, tongue and jaw. When viewed in terms of this relation, vowel and tonal targets appear to be completely independent.

Keywords: tones, vowels, speech production models, Chinese, EMA, coarticulation

1. Introduction

Models of speech production generally assume that the glottal source and the supra-glottal vocal tract filter are independent (Fant 1960; Stevens 1998) – an assumption implicit in models of syllable structure in which vowel quality and tone are independent (e.g. Yip 2002; Duanmu 2007; Gao 2009). In contrast, traditional Chinese phonology divides the syllable into two non-decomposable parts: an 'initial' (*shēngmǔ*) and a 'final' (*yùnmǔ*) (e.g., Chao 1968). The 'initial' is the first consonant of a syllable. The 'final' includes the nuclear vowel, tone and optional coda into a single unit. More holistic supra-phonemic units, such as the finals of traditional Chinese phonology, are consistent as well with more contemporary exemplar-based models that posit word-specific phonetics or online abstraction over exemplars of various-sized units (e.g., Bybee 2003; Pierrehumbert 2001).

In this study, we seek to evaluate the status of vowels as units of speech production that are independent from tone. We expect to find, if vowels are independent from tones, consistent vowel targets across different tones modulo any effects of tone-vowel coarticulation. On the other hand, if units of speech production are larger, more holistic complexes, such as words or 'finals' we expect each tone-vowel combination may have unique spatial targets.

To address the question of tone~vowel independence, we conducted an EMA study of natural variation in tongue displacement across tones. Previous data suggest that tongue position varies to some degree with tone height (Erickson et al. 2004; Hoole & Hu 2004; Hu 2004). These studies report EMA data from one or two speakers with a limited number of contexts, tones and repetitions. More data is needed to evaluate the stability of vowel targets across tones.

2. Method

Six native speakers of Mandarin Chinese (3 male) participated. Each speaker produced multiple repetitions of three maximally-dispersed vowels (/i/-/a/-/u/) in labial-initial syllables (/pV/) with each of the four Mandarin tones: 1 'high', 2 'low-high', 3 'low', and 4 'high-low'. Each syllable was produced 12 times by each speaker, generating a corpus of 864 tokens (12 reps x 3 vowels x 4 tones x 6 speakers). Syllables were presented in Pinyin and randomized with fillers.

We used an NDI Wave electromagnetic articulograph system sampling at 100Hz to capture articulatory movement. The NDI wave supports 5D sensors and 6D sensors, which can be used for automatic head correction by a proprietary algorithm. We used only 5D sensors for this experiment, attached to the tongue tip (TT), blade (TB), dorsum (TD), lips, jaw, nasion and mastoids. Acoustic data were recorded simultaneously with a shotgun microphone. Head movements were corrected computationally after data collection with reference to the mastoid and nasion sensors. The post-processed data was rotated so that the origin of the spatial coordinates corresponds to the bite plane (front teeth). We analyzed the spatial location of the lingual and jaw sensors at the vowel target, using the *findgest* algorithm in MVIEW (Tiede, 2010). Vowel targets were determined by a 20% threshold of peak velocity of the TD sensor in the opening movement of the vowel. All results below are based on measurements taken at this landmark.

3. Results

We report the results in three stages. First we report f0 across the tones and vowels. These results indicate consistent tone patterns across vowels. We then analyze tone-conditioned spatial variation found at the vowel target landmark, focusing on the mid-sagittal plane. After considering the vertical and longitudinal dimensions separately, we introduce a second order measure that integrates displacement in these dimensions relative to the jaw sensor.

3.1. F0 contours

Figure 1 shows the average f0 contour for each vowel across tones. F0 was sampled at regular fixed intervals based on 10 percent of total vowel duration for each tone-vowel combination. The raw f0 samples were converted to z–scores within speaker, an effective normalization procedure for tone (Rose 1987), before averaging. Figure 1 demonstrates that our speakers produced f0 trajectories across vowels that are highly consistent with previous findings for Mandarin Chinese (e.g., Howie 1976), including a small effect of intrinsic f0 on static

high (tone 1) and low (tone 3) tones (c.f., Shi and Zhang, 1987).

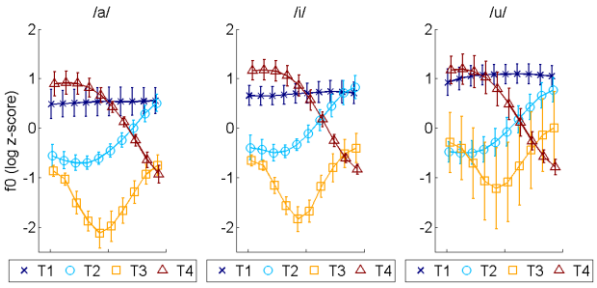

Figure 1: *Average normalized f0, y-axis, plotted by time expressed as a percentages of total vowel duration, x-axis, for each combination of tone (T1-T4) and vowel (/i/-/a/-/u/) in the corpus.*

3.2. **Tongue and jaw sensors across tones**

Figure 2 provides a summary of tongue and jaw position within the mid-sagittal plane across tones and vowels. Each point represents the average spatial position across speakers at the vowel target landmark. Tongue edges are represented as quadratic fits between Tongue Tip (TT), Tongue Blade (TB) and Tongue Dorsum (TD) sensors (within vowel and within tone). Similar tongue and jaw position across tone can be seen for /u/, small differences for /i/ and larger differences for /a/.

Before averaging values across speakers, sensor positions were normalized by z-score. Figure 2 shows average z-scores projected back onto mm units using the group mean and standard deviation for each sensor. This provides a visualization of effect size in mm units. All statistical analyses were conducted on normalized values. Separate repeated measures ANOVAs were conducted on z-scores for each sensor and vowel in the vertical and longitudinal (anterior-posterior) dimensions. Significant effects of tone on vowel position were found in the vertical dimension at the TB sensor for both /i/ [$F(3,15)=5.95$, $p < .01$] and /a/ [$F(3,15)=11.55$, $p < .001$], at the TT sensor for /a/ [$F(3,15)=6.87$, $p < .01$], and at the Jaw sensor for /a/ [$F(3,15)=4.58$, $p < .05$]. The only significant effect of tone on longitudinal position was found for /a/ at the TT sensor [$F(3,15)=4.67$, $p < .05$]. These effects are highlighted with rectangular boxes in Figure 2. There were no significant effects of tone on /u/ in either dimension.

Effects of tone on vowel target position, summarized in Figure 2, were observed in two of the three vowel contexts examined: tones that start low (2 and 3) pattern together in their influence on tongue position. The results for /a/ production are consistent with the findings of Erikson et al. (2004), who observed that the tongue body is lower (and F1 higher) for /a/ produced with a low tone. The broader constellation of effects found for /a/ with tones that start low (lower jaw, lower TB, lower and more posterior TT) points to a common physiological explanation. Reduction of vocal fold tension for low tones can be achieved by lowering the larynx (Honda et al. 1999; Moisik et al. 2014), which could pull the jaw down (Honda 1995). Jaw movement in the opening phase of vowels involves both a rotational component, whereby the jaw rotates around a terminal hinge, the temporomandibal joint, and vertical and horizontal translations of that axis (Edwards and Harris, 1990). The pattern of effects for /a/ is as expected if the effect of tone on vowel position is mediated by the rotational component of jaw movement. Since the jaw lowers in an arc-like motion, greater lingual displacement as a function of jaw movement is expected for sensors distal to the temporomandibal joint. On this account, the relative stability of the TD sensor follows from its posterior position. Given a degree difference in jaw rotation, the magnitude of the effect on sensor displacement is proportional to the distance from the terminal hinge. Thus, the more anterior lingual sensors, TT and TB, show larger effects of tone, as expected if driven by rotational jaw movement. Moreover, for the most anterior lingual sensor, TT, lowering goes hand in hand with retraction. This also is expected if the arc-like motion of the jaw is driving the effect of tone on tongue position.

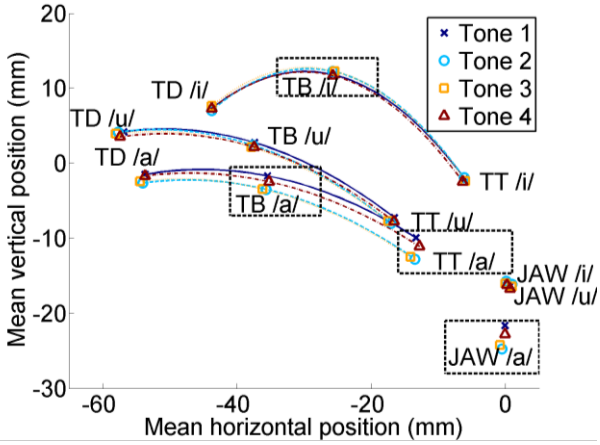

Figure 2: *Mean midsagittal tongue position (12 repetitions, 6 speakers) for Mandarin vowels /i/-/a/-/u/ produced with Tones 1 'high', 2 'low-high', 3 'low', 4 'high-low'. Tongue edges are represented as quadratic fits between Tongue Tip (TT), Tongue Blade (TB) and Tongue Dorsum (TD) sensors (within vowel and within tone). Dotted line rectangles indicate where mean lingual sensor positions differed significantly with tone.*

Accounting for the effect of tone on /i/ production is less straightforward. The only significant difference was found at the TB sensor in the vertical dimension. As with /a/, tones that start low, tone 2 and tone 3, pattern together. However, unlike for /a/, tones 2 and 3 influence /i/ in the opposite direction. For /i/, the TB was higher for tone 2 and tone 3 than for tones that start high, tone 1 and tone 4. Figure 3 zooms in on these differences comparing the effect for /i/ (right) with that found for /a/ (left). Point estimates are mean values of vertical displacement. Error bars represent 95% confidence intervals. Tones 2 and 3 pattern together but they influence /a/ and /i/ in different directions.

The physiological explanation we offered for the effect of tone 2 and 3 on lingual position for /a/ does not generalize straightforwardly to /i/. However, it may be the case that TB is raised for /i/ with low tones to keep the acoustics of /i/ stable across the four tones by countering mechanistic factors. In other words, the same pull of low tones on the jaw may receive lingual compensation for /i/ but not /a/. This may seem odd from the standpoint of articulatory-acoustic dynamics, where the formant values of /i/ are relatively robust to articulatory variation (Stevens, 1989), but the compensation account is reasonable given the vowel space of Mandarin Chinese. Mandarin has only one low monophthong, /a/, but is comparatively crowded in the upper regions of the vowel space, containing /y/, /i/, and /u/ monophthongs (Duanmu 2007). Lingual compensation for tone-induced pull on the jaw for /i/ and /u/ may be driven by increased articulatory precision required to maintain contrast between non-low vowels.

Alternatively, it is possible that low tone production with /i/ involves a different laryngeal mechanism than low tone production with /a/. Recent work on Chinese tones has established two mechanisms of laryngeal articulation engaged in low tone production, one which involves larynx lowering and one which involves larynx raising together with laryngeal constriction (Moisik et al., 2014). Of particular interest is that the stimuli in this study included only words containing the vowel /i/. It is possible that preferences for different mechanisms of low tone production vary with vocalic context, and that larynx raising (for low tones) is more likely in /i/ than in /a/. A complication in applying this mechanistic account to our data involves the fact that tones 2 and 3, the tones that start low, pattern together in their influence on TB height whereas Moisik *et al.* (2014) observe larynx raising only for tone 3.

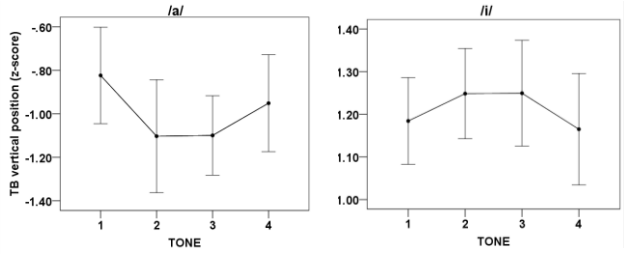

Figure 3: *Mean tongue blade (TB) height (12 repetitions, 6 speakers) for Mandarin vowels* /a/ *(left) and* /i/ *(right) produced with Tones 1 'high', 2 'low-high', 3 'low', 4 'high-low'. Error bars represent 95% confidence intervals.*

3.3. **TB to jaw distance across tones**

To further investigate the relationship between jaw and TB movement, we computed the Euclidean distance between the TB sensor, where opposite effects of tone were found for /a/ and /i/, and the jaw sensor. Unlike the analysis reported above, this measure takes into account changes across tones in both vertical and longitudinal dimensions. Figure 4 summarizes the measurements across speakers. As expected, the TB sensor is farthest away from the jaw sensor for the /u/ target (42.3mm), followed by the /a/ target (40.9mm), and then the /i/ target (38.9mm), which is closest to the jaw. However, in contrast to the analyses of vertical TB displacement reported above, there is a negligible influence of tone on TB-to-jaw distance. A repeated measure ANOVA on TB-to-jaw distance with tone and vowel as independent factors showed a marginal effect of vowel [$F(3,15)=3.83$, $p = .058$] but not tone [$F(3,15) < 1$] and no interaction between tone and vowel [$F(3,15) < 1$].

The null effect of tone on TB-to-jaw distance for /u/ can be expected, since neither the TB sensor nor the jaw sensor was individually influenced by tone. For /a/, we have already seen that both the jaw and the TB sensor were influenced by tone in the vertical dimension. We now see that these parallel movements maintain a fixed distance between TB and jaw sensors. This indicates that the magnitude of jaw sensor displacement is comparable to the magnitude of TB sensor displacement. The result reinforces our view expressed above in the discussion of /a/ that the effect of tone on vowel targets is mediated by jaw movement. The stable TB-to-Jaw distance for /i/ indicates that, here also, the significant effect of tone on vowel height can be attributed to the jaw. This was not apparent from analyses in 3.2 in part because the contribution of the jaw to TB position is divided across vertical and longitudinal dimensions. The TB-to-jaw distance incorporates these into a single measure, bringing out an invariance masked by single dimensional analyses.

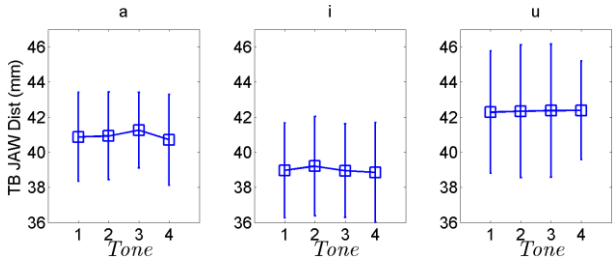

Figure 4: *Mean Euclidean distance between the Tongue Blade (TB) sensor and the Jaw sensor for vowels* /a/-/i/-/u/*, produced with each tone (1-4). Error bars indicate 95% confidence intervals.*

4. Discussion

This study was designed to test whether vowels in Mandarin Chinese have production goals that are independent of tone, as is assumed by modern phonological accounts of Chinese and by most models of speech production, or, alternatively, whether vowels and tones form more integrated composite targets, as in the finals of classical Chinese phonology. There are known physiological linkages between vowel height and vocal fold tension which are mediated by extrinsic muscles acting on the larynx and hyoid bone (Honda 1995). These coarticulatory forces must be taken into account in considering the nature of vowel targets. If vowels and tones have independent production goals then we expect the influence of tone on vowel to fall out from these independent specifications. On the other hand, arbitrary variation, i.e., variation in vowel position that cannot be attributed to the coarticulatory influence of tone, provides support for more holistic speech targets, i.e., tone-vowel inter-dependence.

4.1. **The case for tone-vowel inter-dependence**

Significant effects of tone on lingual position measured at the vowel target were observed for two of the three vowels examined. This result may appear at first blush to support the hypothesis that tones and vowels are inter-dependent, as in the 'finals' of traditional Chinese phonology, more holistic accounts of lexical representation, such as exemplar theory (Pierrehumbert 2002), or other theories that advocate speech production units larger than the vowel, e.g., Fujimura's (1986) icebergs. We offered a partial physiological explanation for why /a/ is lowered for tones that start low, tone 2 and 3. The effect of these tones on /i/ was in the opposite direction. For /i/, TB was higher for tones 2 and 3. We speculated on possible mechanistic (Moisik et al., 2014) and functional accounts for this pattern. However, with respect to the question of tone-vowel independence, maintenance of small but systematic differences in vowel target as a function of tone supports an integrated representational hypothesis. In the absence of a model that can account for why low tones lead to higher TB for /i/ and lower TB for /a/, the pattern appears arbitrary and, as such, supports the inter-dependence view of tone-vowel relations. TB height may vary across /i1/, /i2/, /i3/, and /i4/ in our data because each of these 'finals' are independent units of speech production or because /pi1/, /pi2/, /pi3/, and /pi4/ are all different words of Chinese.

4.2. **The case for tone-vowel independence**

In contrast to the case for tone-vowel *inter-dependence* exposed above, we believe that the data also can be viewed as unequivocally supporting tone-vowel *independence*. However, this interpretation of the data requires that we either focus on certain areas of the tongue, e.g., the stable TD sensor, or that we pursue an alternative expression of vowel targets.

Vowel targets are typically considered to be positions in space, corresponding, for example, to the static images of x-rays (e.g., Stevens and House, 1955). Although details of specific models vary, we take the standard view of vowel targets to involve the position of the tongue relative to the palate. This can be expressed in terms of constriction location and degree, as in Task Dynamics (Saltzman and Munhall, 1989) or as a fixed target in space with quantifiable dimensions (Guenther, 1995). Although there are important differences between these models of vowel targets, they have in common that the production goal is not expressed in terms of the relation between active articulators. Rather, the production goals are expressed independently of the coordinative structures that may achieve them. On this view, to see the data as supporting tone-vowel independence requires focusing on a specific portion of the tongue. Only the TD sensor remains stable across tones. On the view that the entire surface of the tongue contributes to the vowel target, our data instead indicates that vowel targets in Chinese vary with tones in ways that are not fully predictable from physiological constraints on coarticulation, at least not according to our current understanding.

Besides restricting our definition of vowel target to the TD sensor, there is an alternative expression of vowel targets that permits an unequivocal interpretation of the data in terms of tone-vowel independence. This alternative is that it is the *relation* between articulators that serves to dictate production goals for these vowels. Seen through the lens of a relative notion of vowel target, our data provide strong support for tone-vowel independence. The relationship between tongue position and jaw position remained stable across tones, even as sensor position varied, e.g. at the TB sensor for /a/ and /i/. As a consequence, the Euclidean distance between the TB sensor and the jaw differed for phonologically distinct vowels, /a/, /i/, and /u/, but remained constant across tones. It is therefore possible to characterize the three Chinese vowels in this study in terms of a single dimension, the relationship between lingual and jaw position. On this view, spatial variation in lingual position need not disturb achievement of vowel targets, as long as the jaw is free to co-vary in accordance with a vowel-specific tongue-jaw relation.

Of the two perspectives on the data that permit interpretations in terms of tone-vowel independence, it is not yet clear which (TB-to-jaw distance or TD position) is more important for vowel targets. We leave this question to future study.

5. Acknowledgements

We would like to thank our six Mandarin participants as well as Chong Han, Jia Ying, and Yuan Ma for help recruiting them. We are also grateful for preliminary discussions of the data with Allard Jongman, Joan Sereno, San Duanmu, Cathi Best and Denis Burnham. Research was funded by a grant from the MARCS Institute 'Time course of tone perception and production' to Jason A. Shaw, Michael Tyler, Michael Proctor, Donald Derrick and Chong Han.

6. References

Bybee, J. (2003). *Phonology and language use* (Vol. 94). Cambridge University Press.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley: University of California Press.

Duanmu, San. (2007). *The phonology of standard Chinese*. Oxford; New York: Oxford University Press.

Erickson, Donna, R. Iwata, M. Endo & A. Fujino. (2004). Effect of tone height on jaw and tongue articulation in Mandarin Chinese. In *Proc. Intl. symposium on tonal aspects of languages*. Beijing: ISCA.

Fant, Gunnar. (1960). *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. Gravenhage: Mouton.

Fujimura, Osamu. (1986). Relative invariance of articulatory movements: an iceberg model. In J. S. Perkell & D. Klatt (ed.), *Invariance and variability in speech processes*, 226–242. Hillsdale, NJ & London: Lawrence Erlbaum Associates.

Gao, Man. (2009). Gestural coordination among vowel, consonant and tone gestures in Mandarin Chinese. *Chinese Journal of Phonetics* 2. 43–50.

Guenther, Frank H. (1995). Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural Network Model of Speech Production. *Psychological Review 102*, 594-621.

Honda, K. (1995). Laryngeal and extra-laryngeal mechanisms of F0 control. *Producing speech: contemporary issues*, 215-232.

Honda, Kiyoshi, Hiroyuki Hirai, Shinobu Masaki & Yasuhiro Shimada. (1999). Role of vertical larynx movement and cervical lordosis in F0 control. *Language and Speech* 42(4). 401–411.

Hoole, Phil & Fang Hu. (2004). Tone-vowel interaction in standard Chinese. In *Proc. Intl. symposium on tonal aspects of languages*. Beijing: ISCA.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones* (No. 6). Cambridge University Press.

Hu, Fang. (2004). Tonal Effect on Vowel Articulation in a Tone Language. In *Proc. Intl. symposium on tonal aspects of languages*. Beijing: ISCA.

Moisik, Scott R., Hua Lin & John H. Esling (2014). A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS) . *Journal of the International Phonetic Association*, 44(1): 21-58.

Pierrehumbert, Janet B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In Joan L. Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 137–157. John Benjamins Publishing Company.

Pierrehumbert, Janet B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (ed.), *Laboratory Phonology 7*, 101–139. Berlin: Mouton de Gruyter.

Rose, P. (1987). Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Communication*, *6*(4), 343-352.

Saltzman, E., & Munhall, K.G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*, 333-382.

Shi, B., & Zhang, J. (1987). Vowel intrinsic pitch in Standard Chinese. In *Proceedings of the 11th international congress of phonetic sciences* (pp. 142-145).

Stevens, Kenneth N, & House, Arthur S. (1955). Development of a quantitative description of vowel articulation. *The Journal of the Acoustical Society of America, 27*(3), 484-493.

Stevens, Kenneth. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 3-45.

Stevens, Kenneth. (1998). *Acoustic phonetics*. Cambridge: MIT Press.

Tiede, M. (2010). MVIEW: Multi-channel visualization application for displaying dynamic sensor movement.

Yip, Moira. (2002). *Tone*. Cambridge: Cambridge University Press.