TAIWANESE MANDARIN SIBILANT CONTRASTS INVESTIGATED USING COREGISTERED EMA AND ULTRASOUND

Mark K. Tiede[1], Wei-Rong Chen[1], D. H. Whalen[2,1,3]

[1]Haskins Laboratories; [2]City University of New York Graduate Center; [3]Yale University
tiede@haskins.yale.edu; wei-rong.chen@yale.edu; dwhalen@gc.cuny.edu

ABSTRACT

Standard Chinese distinguishes a three-way place distinction among sibilants: (Denti)-Alveolar /s/, 'Retroflex' (Post-Alveolar) /ʂ/, and (Alveolo)-Palatal /ɕ/. While Taiwanese Mandarin generally preserves the standard consonant inventory, previous studies have described its retroflex coronals as being partially merged with alveolars, with higher acoustic center-of-gravity values for retroflex sibilants indicating a more forward place of articulation relative to comparable values for Beijing Mandarin; however these are to date unsupported by kinematic measurements.

Here we examine the articulation of these sounds using electromagnetic articulometry (EMA). Tongue tip and parasagittal blade sensor elevation angles are compared to a reference /s/ position. Additional sensors placed midsagittaly on the tongue blade and dorsum give an index of retroflexion. Concurrently collected ultrasound data, coregistered through reference sensors on the head and probe, provide corresponding midsagittal tongue contours. Results show that although acoustic differences between /s/ and /ʂ/ are small, tongue posture in these sounds is systematically different.

Keywords: Mandarin, sibilants, EMA, Ultrasound, HOCUS.

1. INTRODUCTION

Traditional Chinese phonology classifies the voiceless coronal sibilant fricatives of Standard Chinese (e.g. Beijing Mandarin) into three groups, distinguished by place of articulation. Although subject to dialectal variation, these are broadly described as Alveolar or Denti-Alveolar (/s/), 'Retroflex' or Post-Alveolar (/ʂ/), and Palatal or Alveolo-Palatal (/ɕ/) [5,8]. These sounds also contrast with aspirated and unaspirated affricates produced at the same places of articulation, and in some dialects, including Taiwanese Mandarin, a voiced retroflex fricative (/ʐ/). Because of limited articulatory studies, precise classification of the 'Retroflex' sibilants in particular is controversial, with some authors terming these "laminal post-alveolars" [12], or "apical post-alveolars" [13], or true retroflex articulations [8].

Taiwanese Mandarin (TM) generally preserves the consonant inventory of Standard Chinese (SC). However, previous studies have suggested that retroflex sibilants in TM are gradually losing their distinctiveness from alveolars, with neutralization conditioned to some extent by register and sociolinguistic factors, and with this trend more advanced in the southern districts of Taiwan [10, 3, 14]. A recent acoustic study comparing spectral center-of-gravity (COG) measures showed that although the alveolar-retroflex contrast is still maintained in TM, the retroflex sibilants have a higher COG than their counterparts in SC, indicating a more forward and less distinctive place of articulation [4]. This is consistent with a subsequent combined EMA/palatographic study of TM speakers showing alveolar rather than post-alveolar place constriction for the retroflex sounds [6].

In this work we examine these contrasts as produced by native speakers of TM using a novel approach that coregisters electromagnetic articulometry (EMA) and concurrently recorded ultrasound (US). EMA sensors placed on the head and the US probe support alignment of both sensor positions and extracted tongue contours with vocal tract hard structure. The methods are complimentary, with EMA sensors providing spatial location and angular orientation from fixed points on the anterior tongue and other speech articulators, and US contours imaging the continuous midsagittal tongue profile. By exploiting this co-collection approach, we gain unprecedented access to data bearing on whether the merging alveolar-retroflex contrast in TM represents a shift in articulatory posture or true neutralization.

2. METHODS

2.1. Stimuli

TM coronal fricatives and their matching aspirated and unaspirated affricates as shown in Table 1 were elicited in a range of words that varied the following vowel and tone (28 different contexts in total). These were presented to participants in two forms. The first was within a consistent carrier sentence: 這個__字 (Pinyin [zhe ge __ zi]; "This is a word of __"). The second was within sentences containing multiple instances of different sibilant fricatives; e.g. 咱們怎

麼買了紫色的榨菜 ([za2 men5 zen3 me5 mai3 le5 zi3 se4 de5 zha4 cai4]; "How did we buy a purple pickled mustard"). The motivation was to contrast the more formal *Carrier* context with the more natural speech of the *Sentence* context. Stimuli were presented as sentences to participants using Chinese characters displayed on a computer screen.

Table 1: Elicited sibilant contrasts; (Pinyin)

Manner/Place	Alveolar	Retroflex	Palatal
Fricative	/s/ (s)	/ʂ/ (sh)	/ɕ/ (x)
Voiced Fric.	-	/ʐ/ (r)	-
Affricate	/ts/ (z)	/tʂ/ (zh)	/tɕ/ (j)
Asp. Affr.	/tsʰ/ (c)	/tʂʰ/ (ch)	/tɕʰ/ (q)

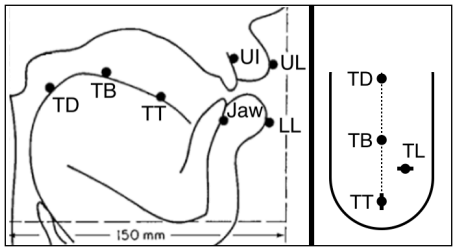
2.2. Participants

One male and two female native speakers of TM have participated in this project to date. M01, aged 41, is from the Northern city of Taipei, and has lived three years in the U.S. F02, aged 53, is from the Northern city of Keelung, and has lived in the U.S. for 20 years. F03, aged 44, is from the Southern city of Tainan, and has spent less than one year in the U.S. Each self-reported normal hearing with no speech deficits, signed informed consent for the experimental protocol approved by the Yale IRB, and were paid for their participation.

2.3. Electromagnetic Articulometry

15 EMA data channels were recorded at 250 Hz using the AG501 system (Carstens), each recording sensor 3D spatial position and angular orientation. Synchronized audio was recorded at 48 kHz through a directional microphone placed ~1 m from the participant's mouth. Reference sensors were placed on the left and right mastoids and the medial upper incisors to correct for head movement. Three more located on the US probe holder provided correction of probe displacement relative to the head. Data sensors were affixed with cyanoacrylate midsagittally to the tongue dorsum (TD, located at the projection of the incisors down to the fully distended tongue); tongue tip (TT, located ~1 cm posterior to the apex); tongue blade (TB, midway between TD and TT); lower medial incisors (JAW); and the upper and lower lip at the vermillion border (UL, LL). Additional sensors were placed parasagittally on the left tongue blade (TL), the left lower canine (JAWL), and the right mouth corner (LC). Figure 1 illustrates sensor layout. TT and TL were placed such that the azimuth defining elevation angle orientation was ~0° (for positive tongue tip angle encoding retroflexion) and ~90° (for positive left blade angle encoding extent of medial grooving) respectively.

Figure 1: Sensor Layout: midsagittal profile and tongue. TT and TL bars show azimuth defining orientation of elevation angles.



2.4. Ultrasound

Ultrasound imaging was performed using the Acuson X300 system (Siemens) with a C8-5 probe at a 30 Hz frame rate, and depth of field adapted to each speaker. Streamed video was recorded at 60 fps with synchronized audio using an image capture device (AverMedia). The probe was aligned for midsagittal imaging and stabilized using the head mount system described in [7].

2.5. Experiment

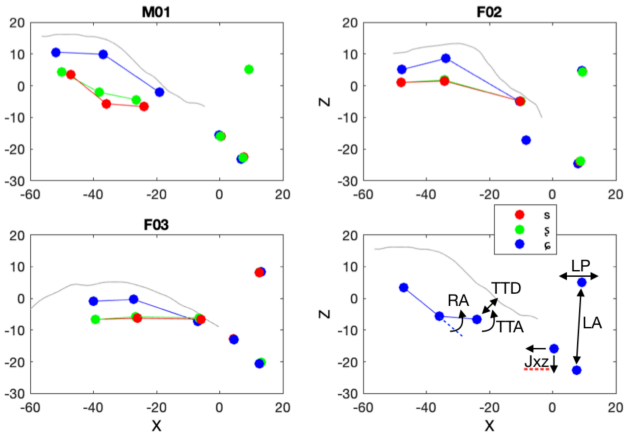
Each experiment session had four parts. First, the stimulus material was recorded in an *audio-only* condition, resulting in 246 sibilant productions. A native speaker of TM monitored recording during each session, and any errorful productions were immediately repeated. Next, the sensors and ultrasound probe were emplaced and their operation validated, following which reference trials were collected to establish each speaker's occlusal plane, the midsagittal outline of the palatal vault, and the location of the probe relative to the head with the jaw clenched. During the main (*EMA*) part of the session, the stimulus material was recorded in pseudo-randomized order within alternating blocks of carrier and sentence contexts, resulting in 492 sibilant productions. The final part of each session consisted of a calibration trial to determine the mapping between the EMA and ultrasound coordinate systems: a sensor was placed at three locations on the US probe surface so as to be visible within the US image, and these reference points, in conjunction with the sensors mounted on the probe holder, established the common origin necessary for coregistration.

2.6. Post-processing

EMA sensor trajectories were low-pass filtered at 20 Hz, corrected for head movement and aligned to each speaker's occlusal plane using the head references.

Forced alignment (P2FA) [17] was used, hand adjusted as necessary in Praat [2], to establish sibilant intervals on recorded trial audio. Temporal alignment of the EMA and US data streams was accomplished using cross-correlation between their common audio. Tongue contours were extracted using [15], converted from pixels to 3D mm, and mapped to the head-corrected EMA coordinate system by relating the calibration position of the probe to its position at the measurement point [16].

Figure 2: Mean midsagittal sensor positions by speaker and sibilant at TT velocity minimum; EMA measures illustrated in lower right panel.



2.7. Analysis

A measurement sampling offset was determined within each EMA sibilant instance by the tangential velocity minimum of the TT sensor. The following measures were obtained from EMA sensor positions at that point (see Fig. 2): TTA (TT elevation angle); TTD (anterio-posterior position of the minimum TT distance to the palate); TLA (TL elevation angle); RA (tongue retroflexion angle determined by TDTB:TBTT); LP (lip protrusion, ULx); LA (lip aperture); LS (lip spreading, LCy); JX and JZ (jaw position). Each of these was converted to z-scores by speaker for analysis.

Following [9] acoustic COG measures were obtained using a multitaper method over a 50 ms window, resulting in measurements of spectral mean (L1), skewness and kurtosis. Linear mixed models to assess effects of context, place and manner on these measures were computed using *lme4* [1] with probabilities assessed using the *lmerTest* [11] package. Log-likelihood comparisons were used to assess whether interaction terms and random slopes by speaker were supported. Significance of model fixed effects was assessed using estimates of the regression coefficients divided by their standard errors (a *t* test), with degrees of freedom based on the Satterthwaite approximation. Significant results are indicated using the p < .001 ***, p < .01 **, p < .05 *, and p < .10 • convention.

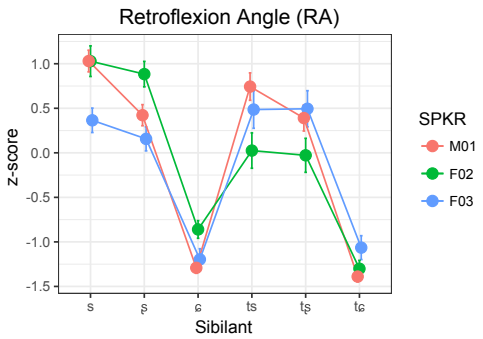
3. RESULTS

3.1. Acoustic

Sibilants require precise control of narrow coronal constrictions that are potentially subject to perturbation by sensors glued to the tongue. To test this possibility, we used fixed effects of session (*Audio-only* vs. *EMA*), context (*Word* vs. *Sentence*), place (*avleolar*, *retroflex*, *palatal*) and +/- affricate to predict sibilant spectral mean (L1), with random slopes and intercepts by speaker (voiced /ʐ/ was excluded and aspiration ignored for this analysis). Interactions were not justified by model comparisons. Results showed a slight marginal increase in L1 as an effect of *Audio* vs. *EMA* ($t = 3.7$ •).

Place showed a hierarchy in L1 frequency with *alveolar* > *retroflex* > *palatal*, though only the palatal contrast was significant ($t = -14.2$ **). Both affricates ($t = -13.0$ ***) and sentence context productions ($t = -5.1$ ***) were systematically lower in L1. A subset of the data evaluating only non-affricated productions found the same general pattern of results for L1, but with no effect of *Audio* vs. *EMA* at all ($t = -0.1$ n.s.). Skewness showed a difference in place (pal > alv, ret; $t = 6.2$ **), and decreased kurtosis was seen in the sentence context ($t = -2.7$ **). Separate models on non-affricates by speaker showed the same L1 hierarchy (*alv* > *ret* > *pal*) for each, but only L1 contrasted alveolar from retroflex significantly ($p < 0.05$, Tukey HSD).

Figure 3: Retroflexion Angle by speaker, measured CCW between extensions of TD:TB and TB:TT.



3.2. Articulatory

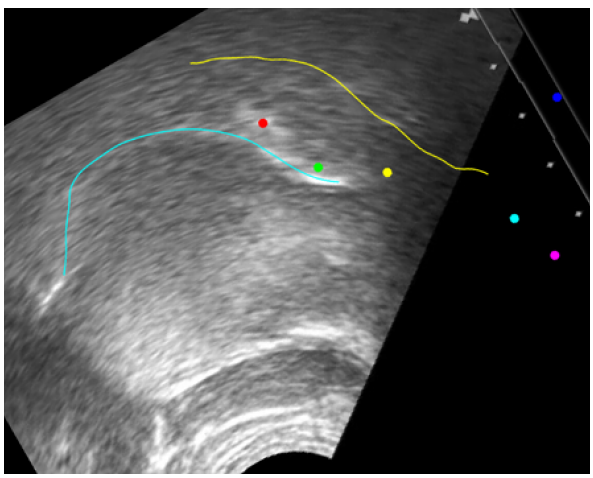
The mean positions of sensors at the TT velocity minimum measurement position are shown in Figure 2, and by-speaker comparisons of the retroflexion angle RA for each sibilant are shown in Figure 3. To evaluate the fixed effects of context (*Word* vs. *Sentence*), place (*alv*, *ret*, *pal*) and +/- affricate on the

articulatory measures described above we used a model with these main effects, with interactions and random slopes for place by speaker included where justified by model comparison.

An effect of place significantly distinguished all three levels for both TTA (*ret* > *alv* > *pal*; Tukey HSD $p < .05$) and RA (*alv* > *ret* > *pal*; $p < .05$), though this interacted for each with manner and context. TTA showed a main effect (increased CCW rotation) for affricate manner (t = 4.9 ***). RA showed reduced retroflexion for affricate manner (t = -7.0 ***) but sentence context increased it (t = 7.0 ***). The parasagittal blade angle TLA showed a negative (doming) effect for palatal place relative to alveolar (t = -3.1 *) and this was enhanced in sentence context (t = -3.2 **).

TTD distinguished retroflex from alveolar place (more posterior; t = -2.7 *), but with sentence context reducing the effect (t = 3.2 **). LP was greater for both palatal (t = 2.9 **) and retroflex (t = 2.3 *) place relative to alveolar. LA was significantly larger for palatal place (t = 4.1 ***) but reduced for sentence production overall (t = -14.8 ***). Lip spreading (LS) was increased for palatal place in affricates (t = 2.1 *) but reduced in sentence production overall (t = -7.4 ***). Jaw showed anterio-posterior retraction (JX) for sentence context overall (t = -3.0 **), and a marginal effect of lowering (JZ) for retroflex place (t = -1.9 •), though this was reduced in interaction with affricate manner (t = 1.9 *).

Figure 4: Speaker M01 producing alveolar /s/ with EMA aligned and superimposed on the corresponding US frame rotated to the occlusal plane. Note US scatter from the TD and TB sensors. The yellow line shows the EMA palate trace; the light blue line shows the fitted US tongue contour.



4. DISCUSSION

The acoustic results indicate that speakers tolerated EMA sensors without significant changes in their sibilant production, validating this approach for observations of this type. The *Word* vs. *Sentence* production context was effective in eliciting a difference in production formality, which was shown by decreased contrast between sibilant types as measured acoustically and in articulation. Only the male speaker produced a significant difference in spectral COG for the alveolar-retroflex place distinction, and only for the *Word* context; this may reflect his upbringing in northern Taiwan where the distinction continues to be more robust than in the south. However, all three speakers produced significantly distinct articulatory measures for this contrast, though these were also reduced under the less formal Sentence production context.

The tongue tip elevation angle TTA and the tongue shape retroflexion angle RA patterned together but not identically; a regression between paired raw values results in an adjusted R^2 of only 0.21 ($F(1,1474)$ = 382.4 ***). In affricated contexts, these measures were opposed, with decreased RA (straighter tongue body) offset by increased TTA curl. Tongue grooving as measured by the parasagittal blade elevation angle TLA was also decreased in affricates. Tongue tip constriction location (TTD) was more retracted for retroflex place, and this was unaffected by affrication. Overall tongue posture, as shown in Figure 2, clearly contrasts two main shapes, with palatal production curling the blade down and alveolar/retroflex production (see also Figure 4) curling it up. The difference between alveolar and retroflex shapes is relatively small and more a matter of degree, with TTD farther back, RA less CCW, and TTA angles more CCW, for retroflex productions. The contrast is enhanced by a trading relation with lip protrusion (LP), enhanced for retroflex, and lip spreading (LS), which is increased in palatal productions.

5. SUMMARY

This work has demonstrated a novel method for co-registration of concurrently recorded EMA and US data streams and applied it to a study of sibilant contrasts in Taiwanese Mandarin. Results are consistent with previous work showing that the TM alveolar-retroflex distinction is less robust acoustically than in Standard Chinese; however, articulatory tongue posture differences remain significantly distinct. In addition, register differences elicited here in the contrasting *Word* vs. *Sentence* production contexts indicate that speakers continue to produce the contrast robustly when called upon to do so in citation form. While current trends may eventually lead to complete neutralization, TM for now retains its three-way sibilant place distinction.

6. ACKNOWLEDGMENTS

This work was supported by NIH grant DC002717 to Haskins Laboratories. Dolly Goldenberg provided valuable assistance with data collection.

7. REFERENCES

[1] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Statistical Software* 67, 1–48.

[2] Boersma, P. 2002. Praat, a system for doing phonetics by computer. *Glot International*, 5.

[3] Chang, Y. C. 1998. Taiwan Mandarin vowels: An acoustic investigation. *Tsing Hua Journal of Chinese Studies* 28, 255-274.

[4] Chang, Y. H. 2012. *Variability in cross-dialectal production and perception of contrasting phonemes: the case of the alveolar-retroflex contrast in Beijing and Taiwan Mandarin.* Doctoral dissertation, University of Illinois at Urbana-Champaign.

[5] Chao, Y. R. 1968. *A grammar of spoken Chinese*. Berkeley, Los Angeles: Univ. of California Press.

[6] Chen, W. R., Chang, Y. C. 2015. Articulatory targets for coronals in Taiwan Mandarin: A study of EMA, palatography, and linguagraphy. *J. Acoust. Soc. Am.* 137, 2381-2382.

[7] Derrick, D., Carignan, C., Chen, W. R., Shujau, M., Best, C. T. 2018. Three-dimensional printable ultrasound transducer stabilization system. *J. Acoust. Soc. Am.* 144, EL392-EL398.

[8] Duanmu, S. 2007. *The phonology of standard Chinese*. Oxford University Press.

[9] Forrest, K., Weismer, G., Milenkovic, P., Dougall, R. N. 1988. Statistical analysis of word-initial voiceless obstruents: preliminary data. *J. Acoust. Soc. Am.* 84, 115-123.

[10] Kubler, C. C. 1985. The influence of Southern Min on the Mandarin of Taiwan. *Anthropological Linguistics* 27, 156-176.

[11] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2016. lmerTest: Tests in linear mixed effects models. R package version 2.0-32.

[12] Ladefoged, P., Maddieson, I. 1996. The sounds of the world's languages. Oxford: Blackwell.

[13] Lee, W. S., Zee, E. 2003. Standard Chinese (Beijing). *JIPA* 33, 109-112.

[14] Lin, Y. H. 2007. *The sounds of Chinese*. Cambridge University Press.

[15] Tiede, M., Whalen, D. H. 2015. Getcontours: An interactive tongue surface extraction tool. *Proc. Ultrafest VII* Hong Kong. https://github.com/mktiede/GetContours

[16] Whalen, D. H., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-LeHouillier, H., Vatikiotis-Bateson, E., Hailey, D. S. 2005. The Haskins optically corrected ultrasound system (HOCUS). *JSLHR* 48, 543-553.

[17] Yuan, J., Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *J. Acoust. Soc. Am.* 123, 3878.