10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Acoustic-Prosodic and Articulatory Characteristics of the Mandarin Speech Conveying Dominance or Submissiveness

Puyang Geng[1], *Wentao Gu*[1], *Keith Johnson*[2], *and Donna Erickson*[3,4]

[1]School of Chinese Language and Literature, Nanjing Normal University, China
[2]Department of Linguistics, University of California, Berkeley, CA
[3]Haskins Laboratories, CT
[4]Kanazawa Medical University, Japan

gengpuyang6@gmail.com, wtgu@njnu.edu.cn, keithjohnson@berkeley.edu, EricksonDonna2000@gmail.com

Abstract

This study investigated the coding strategy for the speech conveying two opposing attitudes, i.e., dominance and submissiveness, based on the utterances elicited by role-play dialogues. Using an electromagnetic articulography (EMA), we collected audio signals and kinematic data of the articulators (including tongue and lips) from 33 native speakers of Mandarin. For dominant speech, prosodic analysis showed a wider F_0 range, a higher intensity, and a faster speech rate, while articulatory analysis exhibited a wider range of tongue vertical movement, a larger lip protrusion, and a larger lip opening than in submissive speech. Results indicate that both prosody and segmental articulation play roles in encoding dominant/submissive attitudes. Dominant speech is characterized not only with a vocal tract expansion (both horizontally and vertically) which supports the frequency code hypothesis, but also with prosodic intensification and hyper-articulation of tongue, in comparison to submissive speech.

Index Terms: prosody, articulatory movement, dominance, submissiveness, electromagnetic articulography

1. Introduction

Dominance and submissiveness constitute a pair of mutually opposite attitudes that are widely used in social affective speech communication [1], especially when the interlocutors have different status. Here, dominance is an aggressive attitude, whereas submissiveness is a non-aggressive one. An in-depth study of dominant or submissive speech is helpful for a better understanding of human speech interaction, for L2 speech acquisition, and for cross-cultural adaption.

According to the "frequency code hypothesis" (henceforth 'FCH'), humans and other mammals use the frequency cues in their vocal sounds to convey the body-size information, and hence the status of (non-)aggressiveness [2, 3]. For example, a lower fundamental frequency (henceforth, F_0) and an "o-face" (i.e., vocal tract lengthening by lip protrusion) are related to a larger body-size impression, thus associated with aggressiveness or dominance. In contrast, a higher F_0 and a smiling face (i.e., vocal tract shortening by lip retraction) are related to a smaller body-size impression, thus associated with non-aggressiveness or submissiveness.

Despite a general consensus on the FCH, very few experimental studies have been conducted on the speech specifically conveying dominant or submissive attitudes, except some reports of a lower F_0 in dominant speech [4-7], as predicted by the FCH. Also, a higher intensity and a shorter duration (i.e., a faster speech rate) was found in dominant speech than in submissive speech [5-8]. To our knowledge, no articulatory study has been reported on dominant/submissive speech, but some studies on emotional Mandarin speech using the Electromagnetic Articulography (EMA) reported that angry speech, which was closely related to aggression, was characterized with a more prominent jaw opening, leading to a larger vocal tract [9, 10], and with a wider range of tongue vertical movement, resulting in hyper-articulation [10].

To make a systematic investigation into the acoustic-prosodic and articulatory characteristics of the Mandarin speech conveying dominant or submissive attitudes, this study collected the audio and kinematic data from 33 native speakers of Mandarin, using an EMA. Both prosodic measurements (including F_0, intensity, and duration) and articulatory measurements (including tongue movements, lip protrusion, and lip opening) were statistically analyzed.

2. Method

2.1. Participants and materials

Thirty-three native speakers of Mandarin Chinese (15 male and 18 female) born in northern China (including Beijing, Hebei, and the northeastern provinces) were recruited for the experiment. Thirty of them were visiting students at the University of California, Berkeley, and three others were graduate students at Nanjing Normal University, China. At the time of data collection, the participants recruited in the US had resided in Berkeley for less than 6 months, and all of them reported daily use of Mandarin.

The mean ages of the male and female participants were 23 and 25 years, respectively. The mean heights of the male and female participants were 176.7cm (SD: 4.7cm) and 164.7cm (SD: 4.6cm), respectively. Height was controlled because body size is positively correlated with vocal fold length and vocal tract size. No participant had a reported history of speech or hearing disorders. All of them were reasonably remunerated for their participation.

Twelve target sentences composed of 4-16 syllables were designed. The sentences were neutral in their literal meaning, but they could be expressed in opposite attitudes (i.e., dominant or submissive) when embedded in different contexts.

Speech data were collected using a paradigm of role-play elicitation. For each target sentence, two scenarios were

10.21437/SpeechProsody.2020-87

designed to elicit two opposite attitudes. Each scenario contained a dialogue of 4-8 turns, with the target sentence always occurring in the last turn where the intended attitude was to be expressed. An instruction text was also provided to elucidate the dialogue situation including the relationship between the interlocutors. To make a better role-play, three cartoon pictures were presented for each scenario.

2.2. Data recording

Audio data were recorded using a lavalier microphone (AKG-C417) sampled at 44kHz. Articulatory data were recorded using a Northern Digital, Inc. (NDI) Wave EMA. To track the movements of the articulators, sensors were placed in the mid-sagittal plane adhered to the upper and lower lips (UL and LL), and the lower incisor (JW, as it indicates the jaw position). A six-degree-of-freedom reference sensor (REF), placed on the forehead in the mid-sagittal plane, was used to correct head movements and to rotate and translate the data onto an occlusal coordinate plane constructed by the three sensors on a bite plate [11, 12]. Figure 1 shows the locations of the sensors on a mid-sagittal view of the vocal tract. The trajectory of each sensor in the magnetic field was recorded by the EMA system at a sampling rate of 200Hz.

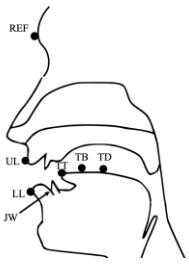

Figure 1: *The mid-sagittal view of the sensor placements in the vocal tract.*

The experiment was conducted in a quiet room in the UC Berkeley Phonlab. Each participant was seated approximately 50cm in front of a computer monitor, and the EMA magnetic field generator was placed about 20cm from the left side of his/her head. A clip microphone was positioned about 25cm away from his/her mouth. The speech stimuli were presented by OpenSesame [13], which was at the same time connected with the EMA data acquisition software Wavefront 2.0.

For each scenario, three cartoon pictures were presented in sequence, at the end of which the participant uttered the target sentence with the preset attitude to match the contextual situation. Thus, altogether 2 (attitude) * 12 (target sentence) * 33 (participant) = 792 utterances were recorded. For each utterance, both audio and articulatory data were collected using the Wavefront 2.0 software.

2.3. Perceptual validation and data processing

In order to test how well the participants produced the intended attitudes, 20 native listeners of Mandarin were asked to judge the attitude of the speaker after hearing a target utterance in a two-alternative force choice (2AFC) task, i.e., to make a choice from 'dominant' and 'submissive.' The overall rate of recognition turned out to be 70.3%. To conduct study on reliable attitudinal speech data, only those utterances with a recognition rate above 75% (i.e., 1.5 times chance level) were adopted for further analysis. As a result, 207 dominant utterances and 206 submissive utterances were analyzed.

Using a Python script [11], all articulatory data were adjusted in relation to the REF sensor position to correct head movements, and then were mapped onto the x-y plane (i.e., the bite plate) defined by the positions of three sensors: the REF sensor, the upper incisor sensor (as the origin of the x-y plane), and the molar sensor placed at the midpoint between the left and right molars. Finally, a robust filtering [14] was conducted to interpolate and smooth the articulatory data.

2.4. Measurements

Audio data were segmented and annotated automatically at the word and phonemic levels using the Montreal Forced Aligner [15], and then were manually corrected by an experienced labeler using Praat [16]. F_0 values were extracted using a short-term autocorrelation algorithm in Praat. After manual correction of gross F_0 errors, the mean and the range of F_0, measured in semitone (st) with a reference of 100 Hz, were calculated for each utterance. Also, the mean and the range of intensity, and the total duration were calculated for each utterance using Praat.

For articulatory measurements, the horizontal (x) and vertical (y) coordinates of tongue (TT, TB, TD), lips (UL, LL) and jaw (JW) sensors were extracted at every sampling time point for each utterance. The horizontal and vertical coordinates of each sensor were then normalized to z-scores among all utterances in each speaker to facilitate inter-speaker comparison. The following procedures were conducted:

(1) To investigate lingual movements, a principal component analysis (PCA) was conducted on all six tongue measurements (TTx, TTy, TBx, TBy, TDx, TDy). The first two components were used to approximate the lingual gestures, with a 90.5% explained variance on average. Roughly, PC1 accounts for the tongue horizontal movement (TH), while PC2 accounts for the tongue vertical movement (TV).

(2) Lip protrusion (LP) was calculated by subtracting the jaw horizontal coordinate from the lower lip horizontal coordinate (i.e. LP = LLx - JWx) for each sampling time point.

(3) Lip opening (LO) was calculated by subtracting the lower lip vertical coordinate from the upper lip vertical coordinate (i.e. LO = ULy - LLy) for each sampling time point.

The means and the ranges of TH, TV, LP, and LO were then calculated for each utterance.

2.5. Statistical analysis

All prosodic parameters (including duration, mean and range of F_0 and intensity) and articulatory parameters (including mean and range of TH, TV, LP, and LO) were statistically analyzed using R [17]. Linear mixed-effects models were conducted to compare these parameters between dominant and submissive utterances, using the "lme4" package [18] and the "lmerTest" package [19]. Tukey *post hoc* tests were then conducted to make pairwise comparisons, using the "lsmeans" package [20]. In each model, attitude (dominant vs. submissive) and gender (male vs. female) were fixed effects, while sentence and participant were taken as random effects.

3. Results

3.1. Prosodic analysis

Table 1 shows the results of linear mixed-effects models (LMMs) on all prosodic parameters. Henceforth, the asterisks

Table 1: *The results of linear mixed-effects models on prosodic parameters.*

Factor	F_0 (st)			
	Mean		Range	
	F	*p*	F	*p*
Attitude	13.37	***	7.78	**
Gender	380.09	***	2.48	0.13
Attitude × Gender	7.25	**	0.00	0.99
	Intensity (dB)			
	Mean		Range	
	F	*p*	F	*p*
Attitude	350.78	***	93.54	***
Gender	0.06	0.81	4.64	*
Attitude × Gender	0.55	0.46	0.06	0.81
	Duration (s)			
	F		*p*	
Attitude	173.90		***	
Gender	0.70		0.41	
Attitude × Gender	4.08		*	

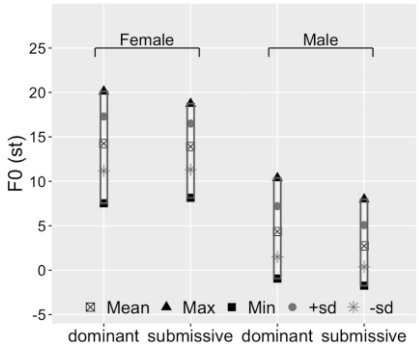

Figure 2: *The average values of the mean, max, and min of F_0 (in semitone).*

in the tables indicate that significant differences exist between the two attitudes (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$).

Figure 2 shows the average values of the mean and range (i.e., max–min) of F_0 for the two attitudes in both genders. The LMM analysis showed a significant interaction effect of "Attitude × Gender" on mean F_0. A Tukey *post hoc* test on the interaction effect showed that dominant speech had a higher mean F_0 than submissive speech only in male ($\beta = 0.81$, $SE = 0.19$, $t = 4.17$, $p < 0.001$), while there was no significant difference in female ($\beta = 0.14$, $SE = 0.16$, $t = 0.85$, $p = 0.4$). There was also a significant main effect of "Attitude" on F_0 range. A Tukey *post hoc* test on the main effect showed a wider F_0 range in dominant speech than in submissive speech ($\beta = 1.11$, $SE = 0.4$, $t = 2.79$, $p = 0.006$). Besides, the main effect of "Gender" was significant on mean F_0 – female had a higher F_0 than male, as expected.

The LMM analysis showed significant main effects of "Attitude" on both mean and range of intensity. Tukey *post hoc* tests on the main effects showed a higher mean intensity ($\beta = 4.32$, $SE = 0.23$, $t = 18.73$, $p < 0.001$) and a wider range of intensity ($\beta = 3.84$, $SE = 0.4$, $t = 9.67$, $p < 0.001$) in dominant speech. No significant interaction effect of "Attitude × Gender" was found on any intensity parameter.

Table 2: *The results of linear mixed-effects models on articulatory parameters.*

Factor	Tongue horizontal movement (TH)			
	Mean		Range	
	F	*p*	F	*p*
Attitude	2.8	0.09	2.45	0.12
Gender	0.59	0.45	0.14	0.71
Attitude × Gender	2.85	0.09	0.02	0.88
	Tongue vertical movement (TV)			
	F	*p*	F	*p*
Attitude	0.14	0.71	14.31	***
Gender	0.18	0.67	0.71	0.41
Attitude × Gender	1.48	0.23	0.47	0.49
	Lip protrusion (LP)			
	F	*p*	F	*p*
Attitude	7.31	**	3.69	0.06
Gender	0.62	0.44	2.16	0.15
Attitude × Gender	2.31	0.13	0.02	0.89
	Lip opening (LO)			
	F	*p*	F	*p*
Attitude	6.67	*	12.93	***
Gender	1.60	0.22	0.23	0.63
Attitude × Gender	0.75	0.39	0.32	0.57

For duration, the LMM analysis showed a significant interaction effect of "Attitude × Gender." A Tukey *post hoc* test of the interaction effect showed a shorter duration in dominant speech than in submissive speech, for both female ($\beta = -0.35$, $SE = 0.03$, $t = -11.97$, $p < 0.001$) and male ($\beta = -0.26$, $SE = 0.03$, $t = -7.51$, $p < 0.001$).

3.2. Articulatory analysis

Table 2 shows the results of linear mixed-effects models (LMMs) on all six articulatory parameters.

For tongue horizontal movement (TH), the LMM analysis showed no significant main or interaction effects.

For tongue vertical movement (TV), the LMM analysis showed a significant main effect of "Attitude" on TV range. A Tukey *post hoc* test on the main effect showed a wider TV range in dominant speech than in submissive speech ($\beta = 0.27$, $SE = 0.07$, $t = 3.78$, $p < 0.001$). No significant interaction effect of "Attitude × Gender" was found on any TV parameter.

For lip protrusion (LP), the LMM analysis showed a significant main effect of "Attitude" on mean LP, and a marginally significant main effect of "Attitude" on LP range. Tukey *post hoc* tests on the main effects showed a larger mean LP ($\beta = 0.16$, $SE = 0.06$, $t = 2.7$, $p = 0.007$) and a wider LP range ($\beta = 0.21$, $SE = 0.11$, $t = 1.92$, $p = 0.055$) in dominant speech than in submissive speech. No significant interaction effect of "Attitude × Gender" was found on any LP parameter. Figure 3 shows the average values of the mean and range of LP for the two attitudes in both genders.

For lip opening (LO), the LMM analysis showed significant main effects of "Attitude" on both mean and range of LO. Tukey *post hoc* tests on the main effects showed a larger mean LO ($\beta = 0.12$, $SE = 0.05$, $t = 2.6$, $p = 0.01$) and a wider range of LO ($\beta = 0.3$, $SE = 0.08$, $t = 3.6$, $p < 0.001$) in

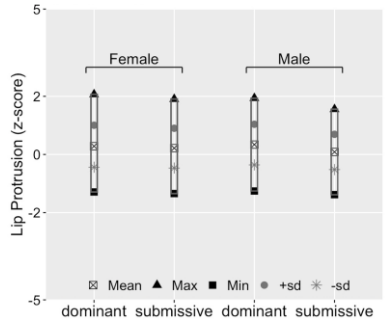

Figure 3: *The average z-scores of the mean, max, and min of lip protrusion (LP).*

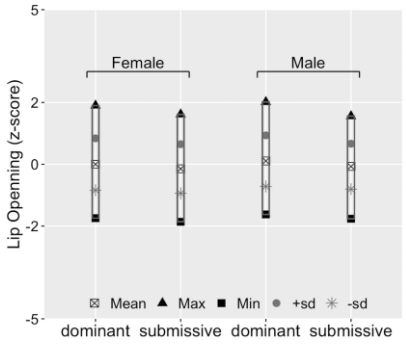

Figure 4: *The average z-scores of the mean, max, and min of lip opening (LO).*

dominant speech than in submissive speech. There was no significant interaction effect of "Attitude × Gender" on any LO parameter. Figure 4 shows the average values of the mean and range of LO for the two attitudes in both genders.

4. Discussion

This study investigated both acoustic-prosodic and articulatory characteristics of the Mandarin speech conveying dominant or submissive attitudes. For prosodic features, dominant speech showed a wider F_0 range, a higher intensity, and a faster speech rate than submissive speech. For articulatory features, dominant speech exhibited a wider range of tongue vertical movement, a larger lip protrusion, and a larger lip opening.

For mean F_0, a gender difference was observed. That is, dominant speech showed a higher F_0 than submissive speech in male, but no significant F_0 difference was found in female. Unlike the findings in previous studies [4-7], this does not coincide with the FCH that predicts a lower F_0 in dominant speech [2, 3], possibly due to differences in speaker or material. On the other hand, the articulatory finding on lip movements, i.e., larger lip protrusion and lip opening (hence a larger vocal tract) in dominant speech than in submissive speech, complies with the FCH, indicating that vocal tract modulation is significant in both anterior-posterior and superior-inferior directions. Taken together, these results suggest that the 'frequency code' can be encoded either in phonation (F_0) or in articulation (vocal tract modulation), but not necessarily both.

Meanwhile, the finding of a wider F_0 range, a higher intensity, and a faster speech rate in dominant speech suggests a kind of "prosodic intensification" when expressing aggressive or dominant attitude, whereas the finding of a wider range of tongue vertical movement in dominant speech than in submissive speech is similar to the previous finding on emotional speech, viz., a "hyper-articulation" of tongue in the speech of high-arousal emotions such as anger [10].

5. Conclusion

The Mandarin speech conveying dominant attitude is characterized by a prosodic intensification (including a wider F_0 range, a higher intensity, and a faster speech rate), a vocal tract expansion in both anterior-posterior and superior-inferior directions which is a support of the frequency code hypothesis, and a hyper-articulation of tongue in the superior-inferior direction, in comparison to the speech conveying submissive attitude.

6. Acknowledgements

This study was funded jointly by a China Scholarship Council scholarship to the first author, and the following two grants: Major Program of the National Social Science Fund of China (13&ZD189), and the project for Jiangsu Higher Institutions' Excellent Innovative Team for Philosophy and Social Sciences (2017STD006). This study was approved by the Committee for the Protection of Human Subjects in University of California, Berkeley (No. 2018-12-11643).

7. References

[1] W. Gu and H. Fujisaki, "Data acquisition and prosodic analysis for Mandarin attitudinal speech," *East Flows the Great River: Festschrift in Honor of Prof. William S-Y. Wang's 80th Birthday*, Hong Kong: City University of Hong Kong Press, pp. 483–500, 2013.

[2] J. J. Ohala, "Ethological theory and the expression of emotion in the voice," in *Proceeding of ICSLP'96*, Philadelphia, PA, USA, vol. 3, pp. 1812–1815, 1996.

[3] J. J. Ohala, "An ethological perspective on common cross-language utilization of F0 of voice," *Phonetica*, vol. 41, no. 1, pp. 1–16, 1984.

[4] H. Mixdorff, A. Hönemann, and A. Rilliard, "Acoustic-prosodic analysis of attitudinal expressions in German," in *Proceedings of the 17th INTERSPEECH*, Dresden, Germany, 2015.

[5] A. Rilliard, D. Erickson, T. Shochi, and J. A. de Moraes, "Social face to face communication: American English attitudinal prosody," in *Proceedings of the 15th INTERSPEECH*, Lyon, France, pp. 1648–1652, 2013.

[6] P. Tang and W. Gu, "Perceptual experiment and acoustic analysis of Chinese attitudes: A preliminary study," in *Proceedings of the 18th ICPhS*, Glasgow, UK, 2015.

[7] K. J. Tusing and J. P. Dillard, "The sounds of dominance: Vocal precursors of perceived dominance during interpersonal influence," *Human Communication Research*, vol. 26, no. 1, pp. 148–171, 2000.

[8] C. D. Aronovitch, "The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker," *The Journal of Social Psychology*, vol. 99, no. 2, pp. 207–220, 1976.

[9] D. Erickson, C. Zhu, S. Kawahara, and A. Suemitsu, "Articulation, acoustics and perception of Mandarin Chinese emotional speech," *Open Linguistics*, vol. 2, no. 1, 2016.

[10] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in *Proceedings of the 6th INTERSPEECH*, Lisbon, Portugal, pp. 497–500, 2005.

[11] K. Johnson and R. Sprouse, "Head correction of point tracking data," submitted.

[12] J. R. Westbury, "On coordinate systems and the representation of articulatory movements," *The Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2271–2273, 1994.

[13] S. Mathôt, D. Schreij, and J. Theeuwes, “OpenSesame: An open-source, graphical experiment builder for the social sciences,” *Behavior Research Methods*, vol. 44, no. 2, pp. 314–324, 2012.
[14] D. Garcia, “Robust smoothing of gridded data in one and higher dimensions with missing values,” *Computational Statistics & Data Analysis*, vol. 54, no. 4, pp. 1167–1178, 2010.
[15] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, *Montreal Forced Aligner [Computer program]*. 2017.
[16] P. Boersma, and D. Weenink, *Praat: Doing Phonetics by Computer [Computer program]*, 2016.
[17] R Core Team, *R: A language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2018.
[18] D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, and H. Singmann, *lme4: Linear mixed-effects models using Eigen and S4. R Package*, 2015.
[19] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, “lmerTest package: Tests in linear mixed effects models,” *Journal of Statistical Software*, vol. 82, no. 13, 2017.
[20] R. Lenth, “Package ‘lsmeans’,” *The American Statistician*, vol. 34, no. 4, pp. 216–221, 2018.