

Articulation, acoustics and perception of Mandarin Chinese emotional speech

Donna Erickson, Chunyue Zhu, Shigeto Kawahara and Atsuo Suemitsu

Abstract

This paper studies articulatory, acoustic and perceptual characteristics of Mandarin Chinese emotional utterances as produced by two speakers, expressing NEUTRAL, ANGRY, SAD and HAPPY emotions. Articulatory patterns were recorded using electromagnetic articulography (EMA), together with acoustic recordings. The acoustic and articulatory analysis revealed that HAPPY and ANGRY were higher pitched, louder, and produced with a more open mouth than NEUTRAL or SAD. SAD is produced with low back tongue dorsum position and HAPPY, with a forward position, and for one speaker, duration was longer for HAPPY and SAD. Moreover, F1 and F2 are more dispersed (i.e. hyperarticulated) in emotional speech than NEUTRAL speech. Perception tests conducted with 18 native listeners suggest that listeners were able to perceive the expressed emotions above chance level. The louder and higher pitched the utterance, the more emotional the speech was perceived. We also explore specific articulatory and acoustic correlates of each type of emotional speech, and how they impact perception.

1. Introduction

1.1. General aims

Emotion is part of our human nature, and it is a subject that is interesting to study from a linguistic perspective. Our everyday experience tells us that speakers can change their speech when they are emotional, which inevitably affects its acoustic realizations. Also, listeners can often, perhaps not always, accurately perceive the emotional state of the speaker. Studying this observation in a scientific manner is one task of modern phonetics.

There is an increasing body of literature on the acoustic characteristics of emotional speech, including acted emotion by professionals and non-professionals and also spontaneous, non-scripted emotional speech (see e.g. Erickson 2005 for a summary). A multitude of different languages have been examined in this research tradition, including Mandarin Chinese (e.g., ChangLiao and Chiang 2004, Gu and Lee 2007, Lin and Fon 2012, Liu and Pell 2012, Yang et al. 2007, Yuan et al. 2002, Wang et al. 2005, Wen et al. 2011, Zhang et al. 2006). However, how speakers articulate emotional speech is less well-studied, with a few exceptions (Erickson et al. 2008, Li et al. 2010, Nguyen et al. 2008). The perception of emotional speech is also less understudied than its acoustics. To fill these gaps, this paper examines the articulation, acoustics, and perception of emotional speech in Mandarin Chinese, in an attempt to further our understanding of speech emotion.

1.2. Previous studies

We start this paper by reviewing some previous phonetic studies on emotional speech, which served as the basis of the current study (see Erickson 2005 for other studies). Erickson et al. (2000) used electromagnetic articulography (EMA) to examine properties of emotional utterances of two American English speakers. They showed that jaw and tongue dorsum position change as a function of the particular emotion, and specifically, the emotion ANGER may involve increased jaw lowering. They also found that emotional speech showed particular acoustic characteristics, realized in terms of F0 and formant structures.

An acoustic and articulatory study on Mandarin Chinese by Erickson et al. (2008) examined a female native speaker, as she was speaking to her friend over a telephone-type connection set up in the lab, recalling past emotional events in her life, including the very SAD story of how her husband was murdered. The acoustic analysis showed that HAPPY syllables were significantly louder, higher in pitch, and shorter in duration than SAD syllables.

Also for SAD, a breathy voice quality was found, as well as lowered lip and jaw, and more tongue tip protrusion compared with HAPPY.

Li et al. (2010) reported on acoustic and EMA recordings of Mandarin Chinese emotional speech (ANGRY, SAD, HAPPY and NEUTRAL, 9 vowels with 111 sentences) for one male speaker. Among their findings were that HAPPY has the highest F0 maximum, followed by ANGRY, and then NEUTRAL/SAD. Articulation also changed as a function of the emotion, such that ANGRY has the highest tongue body position. They also reported increased lip protrusion for SAD as well as for ANGRY. In addition, the study found intonation differences in final boundary tones, with HAPPY having a high rising tone, ANGRY, a high falling tones, and SAD and NEUTRAL, low tones.

Wang et al. (2005) reported that in contrast to neutral speech, the pitch register of HAPPY speech is higher, and the slope of F0 contour of the final syllable of each prosodic word is bigger, especially for the syllable at the end of the sentence.

1.3. The current study

Compared to acoustic examinations of emotional speech, the number of articulatory studies on emotional speech is sparse—at least, more case studies are warranted to advance our understanding of emotional speech. This is most likely due to the challenges of making articulatory recordings such as EMA, which is not available in every phonetics laboratory, and also, the relative newness of such techniques. Our study thus examines the acoustic and articulatory and characteristics of Mandarin as produced by two speakers (male and female), investigating how acoustic and articulatory characteristics change as a function of the speaker's emotion. We also know that not all the utterances were perceived as the emotion intended by the speaker (e.g., Dang et al. 2010, Erickson et al. 2006, Spring et al. 1992). Therefore, we tested what acoustic characteristics allowed listeners to perceive the “right” emotional state.

2. Method

Articulatory and acoustic recordings were made of two Mandarin Chinese speakers producing NEUTRAL, ANGRY, SAD and HAPPY utterances. Perception tests were conducted to see how well the emotions were perceived by listeners.

2.1. Data recording

Articulatory recordings were done using the 3D EMA (Carstens AG500) at the Japan Advanced Institute of Science and Technology. Acoustic signals were simultaneously recorded. Based on the findings reviewed in section 1.2., especially Erickson et al. (2000), this paper reports on the recordings of two EMA sensors: one glued to the lower medial incisors to track jaw motion and another glued to the tongue dorsum (See Figure 1).

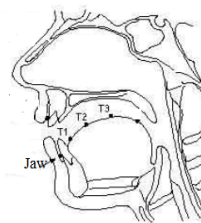


Fig. 1. Placement of sensors on Jaw (Lower Incisor) and Tongue for the EMA experiment. This paper focuses on Jaw and T3 (Tongue Dorsum).

Head movement correction was done using four sensors attached to the upper incisors,

bridge of the nose, left and right mastoid processes behind the ears. The articulatory and acoustic data were digitized at sampling rates of 200 Hz and 16 kHz, respectively. The occlusal plane was estimated using a bite plane with three additional sensors. In post processing, the articulatory data were rotated to the occlusal plane and corrected for head movement using the reference sensors after low-pass filtering at 20 Hz. The lowest vertical position (maximum displacement) of the jaw with respect to the bite plane was located for each target syllable of the utterance using the MATLAB-based software mview (Haskins Laboratories); this measure indicates the amount of the articulator's displacements from the occlusal plane.

Tongue dorsum measurements were made at the time of the lowest vertical jaw position. The horizontal position of the tongue dorsum relative to the upper incisor is indicated as TDx, and low values indicate forward position of the tongue; the vertical position of the tongue dorsum relative to the occlusal plane is indicated as TDz, and large values indicate low tongue dorsum position.

Acoustic measurements of the vowels were made using Praat (Boersma and Weenink 2015). Duration, maximum F0, maximum intensity, F1 and F2 were measured, by creating a 10 ms analysis window centered in the middle of the vowel. Approximately 6 repetitions of each utterance-type were successfully recorded and analyzed.

2.2. Stimuli and analyses

For this paper, we report on 3 Mandarin Chinese phrases: [ba1 na2 ma3] ('Panama'), [ma1 ma ma4 ma3] ('Mother curses the horse'), and [da4 ba3 da4 ba3] ('a lot of'). The vowels in the syllables are all the same, /a/, since jaw displacement varies according to vowel quality. For example, in English, high vowels and mid vowels can differ by 2 mm, and so do mid vowels and low vowels (Menezes and Erickson 2013); similar findings have been reported for Japanese (Erickson and Kawahara 2016; Kawahara et al. submitted). Although no studies we know of have addressed this issue in Mandarin Chinese, we assume that there is a similar relationship between vowel height and jaw displacement. This study thus controlled for the vowel quality in the stimuli. We examined the utterances with the low /a/ vowel, because jaw displacement patterns are most clearly articulated with low vowels, but ongoing work is also looking at utterances with high and mid vowels.

For the current acoustic and articulatory analysis, we focus specifically on the word/phrase final syllable for the following reason. Recent research suggests that utterance-final syllables in Mandarin Chinese are *prominent*, which show increased jaw displacement and increased vowel duration (Erickson, Iwata and Suemitsu (submitted), Iwata et al. 2015). For the speakers in this study, increased jaw displacement for the final syllable is also seen, regardless of the emotion expressed (see Figures 2 and 3 below). Additional motivation for focusing on the last syllable comes from reports by Wang et al. (2005) that the greatest excursion of F0 occurs on the final syllable of each prosodic word and especially for the sentence final syllables.

2.3. Speakers and elicitation

Two native speakers of Mandarin Chinese served as the speakers for the current experiment, one male (C02) and one female (C03), both born and raised in Beijing.

The utterances were spoken with different emotional expressions: NEUTRAL, ANGRY, SAD and HAPPY. The speakers were asked to first speak the NEUTRAL sentences, six randomized repetitions as part of a larger data set; then they were instructed to change their emotional attitude by remembering a situation where they felt very angry, and to speak 6 repetitions of a set of words presented in a randomized order. Then, they were asked to put themselves in a "sad emotional situation" and to speak the sentences, remembering a situation where they

were very sad, and then the same for HAPPY. Between each set of emotions, the speakers were given time to set their moods.

C02 had no experience of acting, but had participated in amateur comedy talk shows in his college days. C02 describes himself as a “cry-baby” i.e., he cries easily when seeing a movie, reading a book, listening to music. So he was able to cry with tears even during the experiment. During the experiment, he imagined himself in the pain of losing his loved ones for SAD, or being angry with someone for something they had said to him for ANGER; for HAPPY, he thought about his son and daughter.

C03 had no experience of acting. When she was doing the experiment, she imagined that she was very mad at someone who did something wrong (ANGER), or she imagined that she would fail to do anything, so what an unfortunate woman she is (SAD).

Each speaker afterwards reported that they experienced physiological changes for each of the emotion sets—especially C02, who was speaking with tears in his eyes for SAD and with his hands shaking and heart beat rising for ANGRY. For this speaker, it took about ten minutes to recover from each emotion, before he could do the next one, and music was played to calm him down. C03 also reported experiencing physiological changes, but not to the same degree.

2.4. Perceptual evaluations

The recorded utterances were randomized to make two perception tests, one for C02 and one for C03, for a total of 72 utterances and 73 utterances, respectively. The stimuli were presented to 18 university Mandarin Chinese students at a Kansai Japanese university. Each listener participated in the two perception tests. The listeners were asked to (a) rate how emotional the utterances were (1 to 5, with “5” as “extremely emotional”, “3” as “emotional”, and “1” as “not emotional”) and (b) identify what emotion they heard—ANGRY, HAPPY, SAD, NEUTRAL, or OTHER EMOTION. Each test was preceded by a practice test of 5 utterances. The tests were presented through headphones in a quiet room.

3. Results

3.1. Perception results

The overall perception test results for speakers C02 and C03 are shown in Tables 1, 2 and 3. Table 1 shows how emotional each utterance was judged by the listeners. Table 2 shows the total number of NEUTRAL, ANGRY, SAD and HAPPY utterances, and how many of these were perceived accurately. Table 3 is a confusion matrix.

Table 1 shows that listeners generally judged emotional utterance as more emotional than the NEUTRAL utterances. All the emotional utterances were rated by listeners with “3” or above (recall that “3” was “emotional”). The highest rated was C02’s expressions of SAD, which were rated as “very emotional”. NEUTRAL was rated as not emotional for both speakers; speaker C02’s HAPPY was rated as not very emotional.

Table 1. Emotional degree (5=very emotional, 3 =emotional, and 1=not emotional).

<i>Speaker</i>	<i>ANGRY</i>	<i>HAPPY</i>	<i>NEUTRAL</i>	<i>SAD</i>
<i>C02</i>	3.3	2.0	1.2	4.2
<i>C03</i>	3.6	3.2	1.3	3.3

Table 2 shows that SAD utterances were best perceived as the speaker intended and HAPPY, the least well perceived.

Table 2. Total number of NEUTRAL, ANGRY, HAPPY and SAD utterances, and their correct and incorrect identification patterns.

<i>Intended Emotion</i>	<i>Correct</i>	<i>Incorrect</i>	<i>Total</i>	<i>%Correct</i>
<i>Neutral</i>	479	133	612	78%
<i>Angry</i>	491	157	648	76%
<i>Happy</i>	423	279	702	60%
<i>Sad</i>	629	19	648	97%

Table 3. Confusion matrix showing results of perception tests.
Perceived Emotion

<i>Intended Emotion</i>	C02				
	ANGRY	HAPPY	NEUTRAL	SAD	OTHER
<i>ANGRY</i>	68%	12%	2%	0%	18%
<i>HAPPY</i>	17%	47%	23%	2%	11%
<i>NEUTRAL</i>	2%	1%	81%	10%	6%
<i>SAD</i>	0%	0%	0%	99%	0%
<i>Intended Emotion</i>	C03				
	ANGRY	HAPPY	NEUTRAL	SAD	OTHER
<i>ANGRY</i>	83%	3%	2%	3%	8%
<i>HAPPY</i>	4%	72%	2%	7%	16%
<i>NEUTRAL</i>	6%	1%	75%	10%	8%
<i>SAD</i>	2%	2%	1%	95%	1%

In Table 3, the accurately-perceived answers are shown in bold. The results show that listeners perceived all expressed emotions better than chance level (1/5 = 20% is chance level). Listeners correct identification rate was highest for SAD for both C02 and C03 (99% and 95%, respectively) and lowest for HAPPY (47% and 72%, respectively). HAPPY for C02 was confused with NEUTRAL (23%) or ANGRY (17%) and for C03, with “other” (16%).

3.2. Acoustic and articulatory results

3.2.1. Correlation analyses on perceived “emotional-ness”

In order to understand what phonetic characteristics contribute to an utterance being perceived as emotional, a correlation analysis was run between the phonetic characteristics of each utterance and the emotional degree, i.e., how emotional the speech was rated by the listeners. The results are shown in Table 4.

Table 4. Results of the correlation analyses between articulatory and acoustic measures and the judged degrees of “emotional-ness”.

	<i>Jaw</i>	<i>TDx</i>	<i>TDz</i>	<i>F0 MAX</i>	<i>F1</i>	<i>Intensity MAX</i>	<i>Duration</i>
C02							
<i>r</i>	0.823	0.038	0.055	0.596	0.607	0.870	-0.091
<i>p</i>	<i>p<.001</i>	<i>n.s.</i>	<i>n.s.</i>	<i>p<.001</i>	<i>p<.001</i>	<i>p<.001</i>	<i>n.s.</i>
C03							
<i>r</i>	-0.222	-0.753	-0.489	0.685	0.428	0.849	0.786
<i>p</i>	<i>n.s.</i>	<i>p<.001</i>	0.042	<i>p<.001</i>	<i>n.s.</i>	<i>p<.001</i>	<i>p<.001</i>

The results show that for both speakers, higher pitched (F0 Max) and louder (Intensity

Max), the more emotional the speech was judged. For C03, in addition, the longer the syllable, the more emotional, whereas for C02, the more open the jaw/the higher the F1, the more emotional. For C03, in addition, the amount of tongue dorsum (TDx) fronting is a cue for emotion: the more fronted the tongue dorsum, the greater the perceived emotion.

3.2.2. Analyses of specific measures

Next, let us move on to the articulatory characteristics of different types of emotional speech. The overall pattern of mean jaw displacement for each of the emotional expressions is shown in Figures 2 and 3; the larger size bar indicates a greater jaw displacement/mouth opening. One general pattern we observe is that jaw opens the most in the sentence-final syllables, regardless of the emotional type. Another observation is that on the utterance-final syllables, C02 has large jaw opening for ANGRY voice, whereas C03 shows largest jaw opening for HAPPY speech. C02 also has larger jaw opening for HAPPY compared to NEUTRAL and SAD.

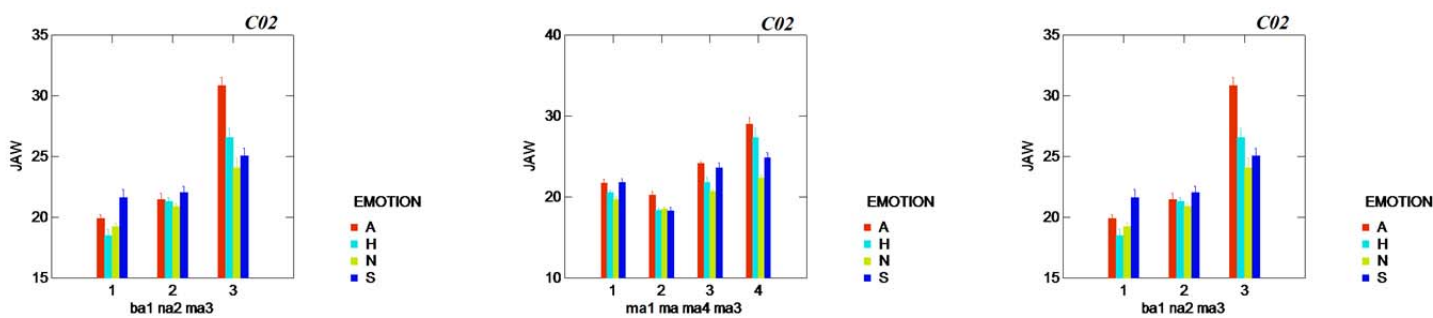


Fig. 2. Amount of jaw displacement by syllable for all utterances that were well-perceived as the emotion intended by speaker C02. (All analyses in this paper concern the emotions that were well-perceived as the emotion intended by speaker.). The y-axis indicates the amount of jaw displacement, so that the larger value indicates a larger mouth opening. Error bars show standard error.

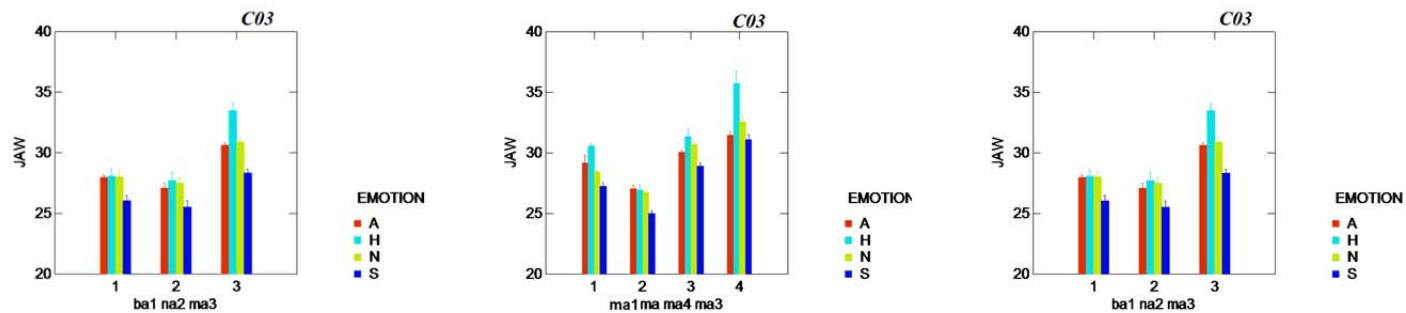


Fig. 3. Bar graph showing amount of jaw displacement by syllable in the word for all utterances that were well-perceived as the emotion intended by C03.

In order to better understand the salient articulatory and acoustic characteristics of each well-perceived emotional expression (NEUTRAL, ANGRY, HAPPY, SAD), Tables 5 and 6 show the mean values (the standard deviations) of the phonetic characteristics of the final syllable: maximum jaw displacement, tongue dorsum; also Maximum F0, Maximum intensity, duration and F1 and F2.

Table 5. Mean values and standard deviations of phonetic measurements for the final syllables for each emotional category for C02. The jaw and tongue dorsum (TDz) measurements are in terms of mm from the occlusal plane; the tongue dorsum-x (TDx) measurements are such that larger negative numbers indicate tongue position is further back in the mouth.

<i>Emotion</i>	<i>Jaw</i>	<i>TDx</i>	<i>TDz</i>	<i>F0 MAX</i>	<i>F1</i>	<i>F2</i>	<i>Intensity MAX</i>	<i>Vowel Duration</i>
<i>SAD</i>	24.7 (1.3)	-67.7 (1.4)	22.4 (2.0)	192.7 (32.6)	869.9 (160.7)	1532.0 (313.8)	71.5 (5.5)	0.3 (0.1)
<i>ANGRY</i>	29.5 (1.9)	-62.4 (2.0)	19.8 (2.2)	204.6 (36.2)	907.7 (71.5)	1230.0 (59.9)	81.9 (4.5)	0.3 (0.1)
<i>HAPPY</i>	26.7 (2.0)	-60.1 (1.9)	19.8 (1.6)	155.2 (15.8)	911.8 (65.6)	1354.3 (42.1)	73.0 (2.2)	0.3 (0.0)
<i>NEUTRAL</i>	22.9 (1.6)	-62.4 (1.3)	19.4 (1.2)	145.6 (38.6)	749.2 (106.0)	1235.5 (57.9)	68.0 (4.0)	0.3 (0.1)

Table 6. Mean values and standard deviations of phonetic measurements for the final syllables for each emotional category for C03.

<i>EMOTION</i>	<i>Jaw</i>	<i>TDx</i>	<i>TDz</i>	<i>F0 MAX</i>	<i>F1</i>	<i>F2</i>	<i>Intensity MAX</i>	<i>Vowel Duration</i>
<i>SAD</i>	29.5 (1.4)	-55.4 (1.1)	7.8 (1.0)	221.0 (26.8)	775.6 (64.4)	1333.6 (25.0)	81.7 (1.8)	0.34 (0.0)
<i>ANGRY</i>	31.3 (0.9)	-52.5 (1.1)	5.5 (0.9)	252.4 (43.5)	1049.2 (115.3)	1473.3 (44.0)	88.7 (2.1)	0.37 (0.0)
<i>HAPPY</i>	34.8 (2.1)	--	--	275.8 (63.9)	1051.3 (128.7)	1458.7 (33.4)	84.3 (2.3)	0.29 (0.0)
<i>NEUTRAL</i>	32.0 (1.5)	-47.9 (1.2)	4.9 (1.0)	162.0 (38.0)	815.4 (172.9)	1413.3 (33.1)	75.0 (1.9)	0.28 (0.1)

In order to examine whether the emotions differ significantly in terms of their phonetic characteristics, ANOVAs were run with item as the random factor, EMOTION as the independent factor and each of the 8 measured phonetic values as the dependents variables. The results, shown in Tables 7 and 8, demonstrate that for both speakers, each of the phonetic values changes significantly, depending on the emotional expression; the one exception is duration for C02 for whom it does not change as a function of the emotion.

Table 7. Results of ANOVA with EMOTION as the factor and each of the 8 measured phonetic values as the dependents for C02

<i>Measured value</i>	<i>F</i>	<i>p</i>
<i>Jaw</i>	49.349	$p < .001$
<i>TDX</i>	64.005	$p < .001$
<i>TDZ</i>	9.862	$p < .001$
<i>F1</i>	22.602	$p < .001$
<i>F2</i>	29.354	$p < .001$
<i>F0 Max</i>	13.260	$p < .001$
<i>Intensity Max</i>	35.090	$p < .001$
<i>Vowel Duration</i>	0.254	$p = 0.858$

Table 8. Results of ANOVA for C03.

<i>Measured value</i>	<i>F</i>	<i>p</i>
<i>Jaw</i>	34.435	<i>p</i> <.001
<i>TDX</i>	347.000	<i>p</i> <.001
<i>TDZ</i>	29.426	<i>p</i> <.001
<i>F1</i>	23.518	<i>p</i> <.001
<i>F2</i>	30.886	<i>p</i> <.001
<i>F0 Max</i>	18.521	<i>p</i> <.001
<i>Intensity Max</i>	118.107	<i>p</i> <.001
<i>Vowel Duration</i>	8.716	<i>p</i> <.001

A post-hoc pairwise comparison analysis was done, whose results are shown in Tables 9 and 10.¹ For both speakers jaw displacement is significantly different for each emotion for all pairwise comparisons, except for ANGRY vs. NEUTRAL for C03. As for TDx, the tongue horizontal (forward-backward) position is significantly different for each of the emotions, except for ANGRY vs. NEUTRAL for C02. As for TDz, for both speakers, the tongue vertical (up-down) position is significantly different for SAD compared to the other emotions, except that for C03, there is no significant difference for HAPPY vs. SAD. F1 for C02 is significantly different for NEUTRAL vs. ANGRY/HAPPY but not ANGRY vs. HAPPY; for C03, F1 is significantly different for all emotions, except that HAPPY is not different from ANGRY, and that SAD is not different from NEUTRAL. F2 is not significantly different for NEUTRAL vs. ANGRY for C02, and not for ANGRY vs. HAPPY for C03. F0 Max is significantly different for all the comparisons for C02 except for HAPPY vs. NEUTRAL and ANGRY vs. SAD, and for C03, except for ANGRY vs. HAPPY. Intensity Max is significantly different for all emotions for C03, but not for SAD vs. HAPPY for C02. As for duration, C02 shows no significant changes in duration as a function of emotion, whereas C03 does, except for HAPPY vs. NEUTRAL/ANGRY.

Table 9. P-values for pairwise comparisons. The p-values for results of pairwise comparison for C02. SAD formant frequencies could not be measured, because of its heavy creakiness.

	<i>ANGRY-NEUTRAL</i>	<i>SAD-NEUTRAL</i>	<i>HAPPY-NEUTRAL</i>	<i>HAPPY-SAD</i>	<i>ANGRY-SAD</i>	<i>ANGRY-HAPPY</i>
<i>Jaw</i>	<i>p</i> <.001	<i>p</i> <.01	<i>p</i> <.001	<i>p</i> <.05	<i>p</i> <.001	<i>p</i> <.01
<i>TDx</i>	<i>n.s.</i>	<i>p</i> <.001	<i>p</i> <.01	<i>p</i> <.001	<i>p</i> <.001	<i>p</i> <.05
<i>TDz</i>	<i>n.s.</i>	<i>p</i> <.001	<i>n.s.</i>	<i>p</i> <.01	<i>p</i> <.01	<i>n.s.</i>
<i>F1</i>	<i>p</i> <.001	--	<i>n.s.</i>	--	--	<i>n.s.</i>
<i>F2</i>	<i>n.s.</i>	--	<i>n.s.</i>	--	--	<i>p</i> <.001
<i>F0 MAX</i>	<i>p</i> <.001	<i>p</i> <.01	<i>n.s.</i>	<i>p</i> <.01	<i>n.s.</i>	<i>p</i> <.001
<i>Intensity</i>	<i>p</i> <.001	<i>p</i> <.05	<i>p</i> <.001	<i>n.s.</i>	<i>p</i> <.001	<i>p</i> <.001
<i>Duration</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>	<i>n.s.</i>

Table 10. P-values for pairwise comparisons. The p-values of pairwise comparison for C03. TDx and TDz could not be measured for HAPPY due to sensor attachment problems.

¹ Since this analysis involves multiple comparisons, the p-values should be interpreted with care, as it may inflate the Type-1 error.

	ANGRY- NEUTRAL	SAD- NEUTRAL	HAPPY- NEUTRAL	HAPPY- SAD	ANGRY- SAD	ANGRY- HAPPY
<i>Jaw</i>	<i>n.s.</i>	$p < .001$	$p < .05$	$p < .001$	$p < .01$	$p < .001$
<i>TDX</i>	$p < .001$	$p < .001$	--	--	$p < .001$	--
<i>TDZ</i>	<i>n.s.</i>	<i>n.s.</i>	--	--	$p < .001$	--
<i>F1</i>	$p < .001$	1.000	$p < .01$	$p < .001$	$p < .001$	<i>n.s.</i>
<i>F2</i>	$p < .05$	$p < .001$	$p < .01$	$p < .001$	$p < .001$	<i>n.s.</i>
<i>F0 MAX</i>	$p < .001$	$p < .001$	$p < .001$	$p < .05$	$p < .05$	<i>n.s.</i>
<i>Intensity</i>	$p < .001$	$p < .001$	$p < .001$	$p < .01$	$p < .001$	$p < .001$
<i>Duration</i>	$p < .001$	$p < .01$	<i>n.s.</i>	$p < .001$	<i>n.s.</i>	$p < .01$

As an additional way of investigating the characteristics of well-perceived HAPPY, SAD, ANGRY and NEUTRAL speech, we present a set of 4 scatter plots of the measured data. Figure 4 shows the scatter plots for F1 and F2 for both speakers.

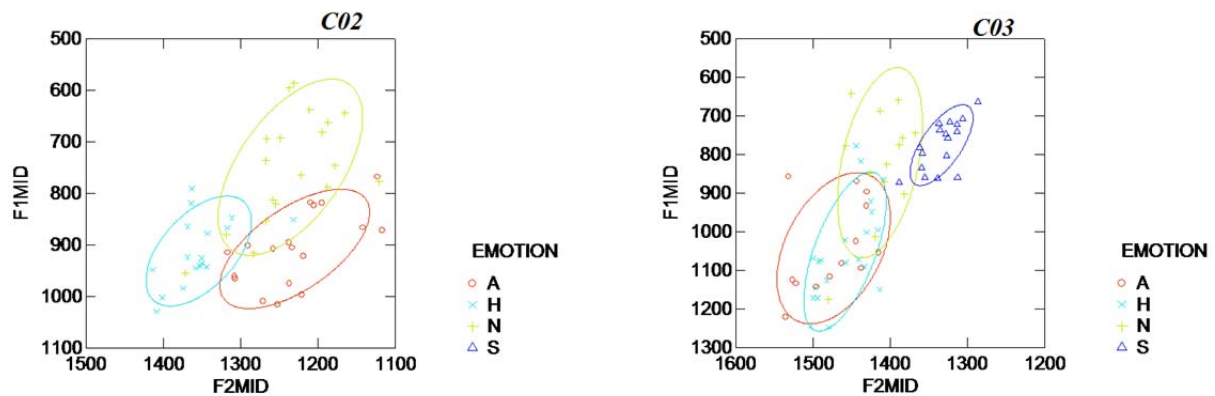


Fig. 4. F1 vs. F2. The graphs show the formant values in terms of the standard phonetic vowel chart (e.g., Ladefoged 2005), such that a high front vowel is at the top left of the graph. The left panel shows values for C02, the right panels, for C03. The colors indicate the different well-perceived emotions: red (ANGRY), turquoise (HAPPY), yellow (NEUTRAL) and blue (SAD). The formant frequencies for SAD for C02 were difficult to measure, since heavy creaky voice occurred frequently on the final syllable, and are not shown here. Ellipses show ellipse sample (ELL, 0.683).

Figure 4 shows that vowel quality of the /a/ changes as a function of the emotion. For both speakers, the /a/ vowel for HAPPY and ANGRY have higher F1 (i.e. the mouth is more open); in terms of F2, HAPPY /a/ for C02 has higher F2 (i.e. the tongue is more fronted) than ANGRY, but about the same degree of frontness for C03. For both speakers, the vowels in NEUTRAL speech tends to be more central, which indicates that emotional speech may involve some sort of hyperarticulation. For C03, SAD is more back and higher.

As for F2, for C02 it is highest for HAPPY and significantly higher than NEUTRAL or ANGRY. For C03, in contrast, F2 is significantly higher for ANGRY/HAPPY (no significant difference) vs. NEUTRAL vs. SAD.

F0 and intensity also change as a function of the emotion. Figure 5 shows scatter plots for C02 and C03, respectively.

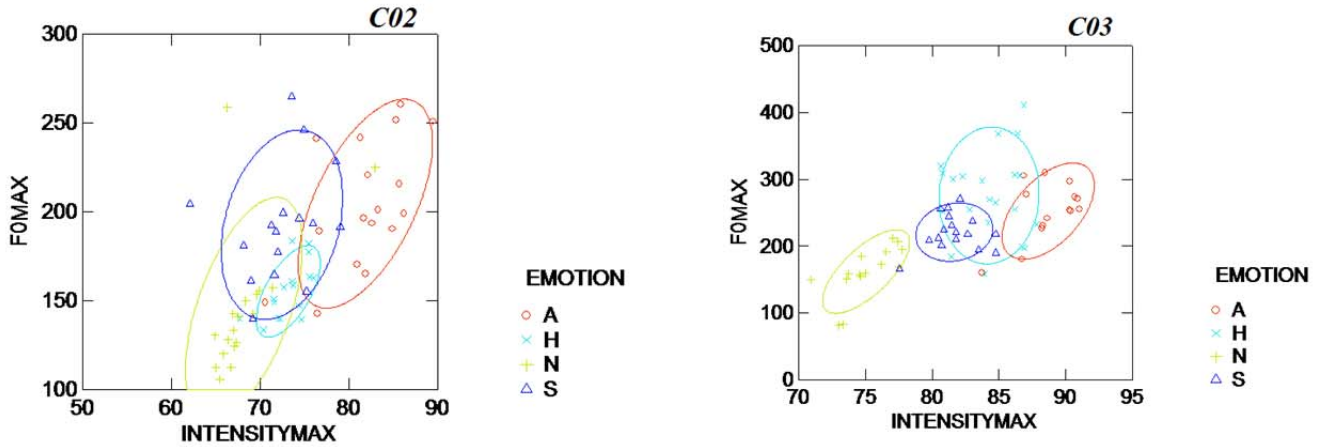


Fig. 5. F0 max in Hz (y-axis) and intensity max in dB (x-axis).

For F0 Max of C02, ANGRY is higher than HAPPY, and SAD is higher than HAPPY/NEUTRAL (no significant difference between HAPPY vs. NEUTRAL). For C03, F0 Max is significantly different for all the emotions, except ANGRY vs. SAD; specifically, HAPPY is higher than ANGRY, then SAD, and then NEUTRAL. As for Intensity Max, for C02, it is significantly different for all the emotions, except HAPPY vs. SAD; specifically ANGRY is louder than HAPPY/SAD, and then NEUTRAL. For C03, Intensity Max is significantly different for all the emotions; specifically ANGRY is louder than HAPPY, then SAD, then NEUTRAL. Overall, in terms of Max F0 and Max Intensity, the emotions tend to be better separated for C03 than for C02.

Figure 6 plots jaw opening against duration.

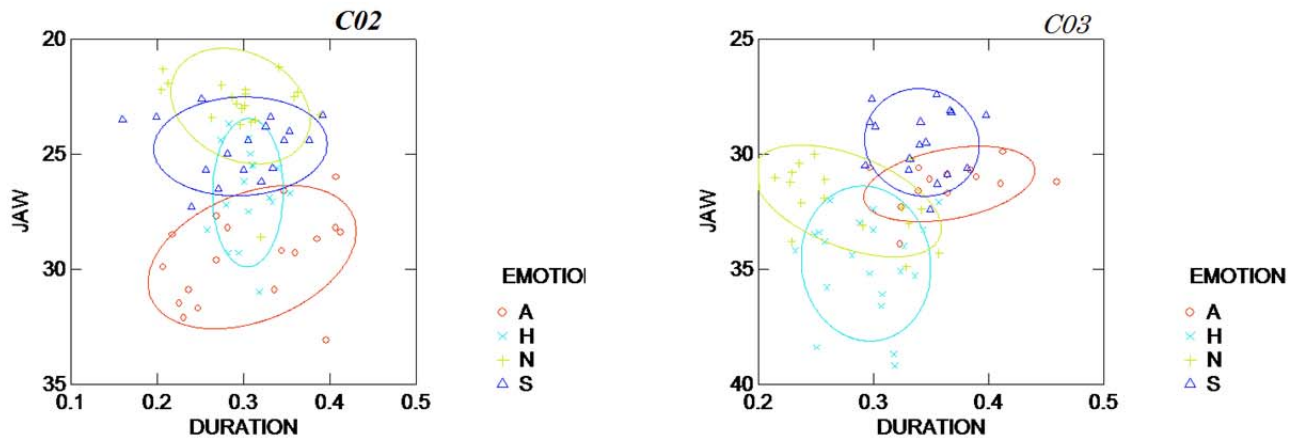


Fig. 6. Jaw displacement in mm with most open jaw at bottom of graph (y-axis) and duration of vowel in ms (x-axis).

For C02, the jaw is significantly different for all the emotions, with significantly larger jaw displacement for ANGRY, then HAPPY, then SAD, and then NEUTRAL; however, duration does not change significantly. For C03, the jaw is significantly different for all the emotions,

except for ANGRY vs. NEUTRAL. HAPPY syllables (not ANGRY), are associated with the most open mouth, then NEUTRAL, and then ANGRY and SAD. With regard to duration, for C03, it is significantly longer for ANGRY/SAD than HAPPY/NEUTRAL. In addition, for C03, there is an interesting relationship between jaw opening and duration as a function of the emotion: for ANGRY, it is mostly vowel duration that tends to increase, while for HAPPY and SAD, it is mostly jaw opening that increases. For NEUTRAL, we see that as the jaw opening increases, duration becomes longer (a similar finding for NEUTRAL was also reported by Iwata et al. 2015 and Erickson et al. 2016).

Finally, looking at Figure 7 for C02, we see that SAD has the lowest, most back TDx position; for both C02 and C03, the TDx position is significantly different for all the emotions, except for ANGRY vs. NEUTRAL for C02. As for the TDz for both speakers, it is significantly lower for SAD than the other emotions.

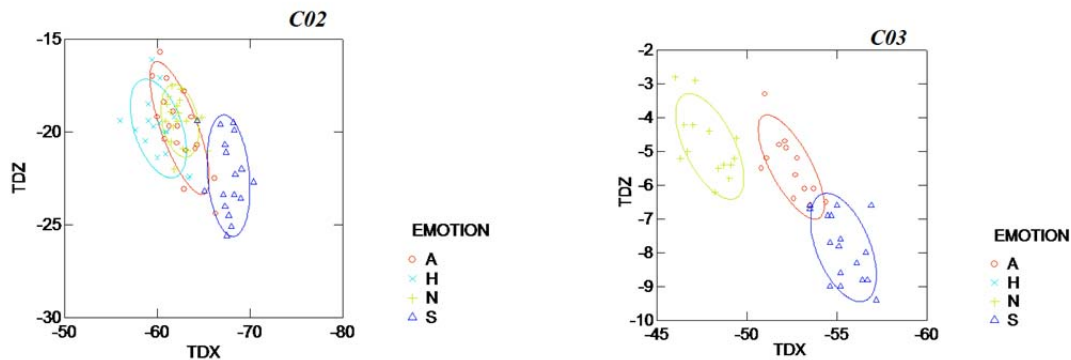


Fig. 7. TDz in mm (y-axis) and TDx in mm (x-axis). The axes are shown such that the speaker is facing left, with the more open mouth and more front tongue at the bottom left of the graph.

4. Discussion

First let us summarize the phonetic characteristics that cue an utterance being heard by listeners as emotional, as reported in Table 4. The results suggest that louder and higher pitched utterances are heard by listeners as “emotional”, a finding also reported by other studies on emotional speech (see e.g., Erickson 2005). Additionally, increased F1 served as a cue to emotion, similarly reported for happy laughing speech by Szameitat et al. (2007). For one speaker (C03), increased duration is a cue to an utterance being heard as emotional, also as reported by e.g. Erickson (2005). In terms of articulation, for one speaker (C02), the more open the jaw, the more listeners heard the utterance as emotional (see also Erickson et al. 2000); also for this speaker, the lower the tongue dorsum, the greater the perceived emotion, especially for SAD.

The second topic concerns whether listeners could perceive the specific emotions intended by the speakers. This study shows that listeners perceived all emotions better than chance. SAD was best perceived, especially for C02, and HAPPY least well. That C02’s SAD was well-perceived as emotional is not surprising, considering that this speaker was actually crying with tears while speaking. That HAPPY in general is not as well perceived as the other emotional expressions is not surprising either, since similar findings have also been reported in the literature (e.g., Erickson 2005, Scherer et al. 2001). More work needs to be done to address why—however, our conjecture based on work with social affective expressions (e.g., Shochi et al. 2009) is that HAPPY is a socially positive expression and as such, is less well marked. ANGRY and SAD are more marked in that they convey information related to survival

and self-protection. Another reason that C02's productions were especially difficult for listeners to recognize as HAPPY may be because the maximum F0 for her HAPPY was rather low, whereas generally HAPPY has a high F0 (e.g., Li et al. 2010).

It is also surprising that C02's ANGRY expressions were not better perceived as emotional, especially since C02 was actually physically shaking with anger during the recordings. This may have to do with the fact that the nine ANGRY utterances that ended without rising intonation were those that were perceived as ANGRY, while those nine ANGRY utterances that were perceived as HAPPY ended with rising intonation. A similar finding about boundary tones and emotions was reported by Li et al. (2010), i.e., ANGRY utterances tend not to have boundary tones, while HAPPY ones had rising ones.

A third topic in this discussion concerns our findings about the acoustic and articulatory characteristics on the word/phrase final syllables of well-perceived emotions (HAPPY, ANGRY, NEUTRAL, SAD). In terms of F0, HAPPY had the highest F0 Max for one of the speakers (C03), similar to that reported by Li et al. (2010) and Wang (2005); however, for the other speaker, C02, ANGRY had the highest F0 Max, and perhaps this is one of the reasons that his ANGRY was sometimes confused with HAPPY, as discussed above. However, for both speakers, ANGRY was the loudest.

In terms of jaw displacement, the largest jaw opening is for ANGRY for C02 (see Erickson et al. 2000 who reported a similar finding), while it is for HAPPY for C03. As for vowel duration, we see for C03 that the negative emotions (SAD and ANGER) have longer durations, as was also reported by Lin and Fon (2012); also by Zhang et al. (2006) for SAD. For C02, however, we do not see this. Concomitant increases in jaw opening and duration have been reported for final (NEUTRAL) syllables by Iwata et al. (2015). It is interesting that for emotional speech, the jaw and duration may, at least for some speakers, work independently. This interplay between emotion, jaw displacement and duration needs to be investigated further.

In terms of tongue dorsum positions, for both speakers SAD has a low back tongue dorsum position and HAPPY the most forward tongue dorsum. It is interesting that tongue positions (Figure 7) do not seem to match the F1-F2 patterns (Figure 4). Prior work with English comparing formant frequencies with TD x-y positions of emphasized vs. non-emphasized syllables shows a strong match with tongue articulation and formants: the more front and high the tongue, the lower the F2 and F1, respectively (Erickson 2002). However, for emotional speech, at least for these two speakers, we do not clearly see this pattern. We do not have an explanation for this discrepancy at present, other than the conjecture that different types of emotional expressions are produced with different types of tongue (as well as lip, and jaw) articulations. It is possible that these articulatory changes lead to changes in voice qualities. This is a topic of future research.

Our data suggest that, in general, the jaw and tongue dorsum work together—when the jaw lowers, the tongue dorsum also lowers; however, for SAD, it seems that the jaw and tongue dorsum may be working independently: for SAD, the tongue dorsum lowers relatively more than the jaw. Moreover, for SAD, the tongue dorsum also moves significantly back. So for SAD, the tongue seems to be moving back and down. A hypothesis, based on previous work with SAD speech suggests that for SAD, the larynx is lowering, and this lowering is pulling the tongue back and down, resulting in a “sad” voice quality to the speech (see Erickson et al. 2006 and Erickson et al. 2008 for more details on this topic). More extensive work is needed in order to understand the relationship between articulation and acoustics in emotional speech.

In actuality, emotions are complex and often there is no single emotion expressed in an utterance (see e.g., Dang et al. 2010). Moreover, the emotional labels used can lead to confusing results. For instance, in this paper, we have analyzed what is often referred to as

“Hot Anger”, in contrast to “Cold Anger”. The characteristics of these two types of angers are different; for instance, the former tends to be high-pitched, loud (e.g., Scherer 1989), and also increased jaw displacement (Erickson et al. 2000), whereas the latter is low-pitched, soft (e.g., Scherer 1989) and decreased jaw displacement (Kim et al. 2014). Also, the term, “Sad”—frequently it refers to a soft, low pitched expression (e.g., Scherer 1989) which is what C03 performed; this contrasts with active grieving sadness (e.g., Erickson et al. 2006, Scherer 1989), which is what C02 performed.

One interesting remaining question is, when experiencing emotion, does the speaker have an acoustic target—i.e., wants to produce a sound such that listeners perceive the emotion? Or, does the emotional experience cause changes in the articulation that result in the acoustic changes? Along these lines, according to Nyugen et al. (2008), spontaneous emotional speech has different glottal characteristics than acted emotional speech. Erickson et al. (2006) found that acted emotions are better perceived than real spontaneous emotion. Erickson et al. (2009) reported that smiling speech was sometimes judged as SAD, and they discussed possible reasons for discontinuities between how a speaker produces the sound, and how it is perceived by listeners. Erickson et al. (2006) reported that there may be a difference in the phonetic characteristics of emotional speech as produced by a speaker in a highly intense emotional situation vs. those characteristics evaluated by listeners as being emotional. These findings may suggest that a highly emotional person may be “inside his/her own world” where the expression of emotion is a personal, cathartic activity. This contrasts with a more “acted” style of emotion, in which the speaker is aware of both him/herself and the other outside person, with the emotional expression more a communicatory activity. With regard to the current two speakers, we might say that the expressions of C03 were well-expressed, well-perceived acted emotions, whereas those of C02, especially for SAD and ANGRY were spontaneous, heart-felt, experienced emotions. The type of spontaneous expressions of emotion by C02 may have no acoustic target *per se*, whereas those of C03 may indeed have acoustic targets, along the lines advocated by Ohala (1994). Much more research is needed on the challenging and complex topics of acted vs. experienced emotions, and the larger topic of acoustic vs. articulatory targets.

5. Summary

This study examined the acoustic and articulatory changes associated with emotions as produced by two speakers of Mandarin Chinese. The results indicate that a speaker’s voice changes in terms of acoustics and articulation when he or she is expressing different emotions. Moreover, listeners are able to make judgments about how emotional the speech is, and what specific emotion the speakers is expressing. In general, emotional speech, and especially ANGRY or HAPPY, tend to be louder and higher pitched; moreover, F1 and F2 are more dispersed (i.e. hyperarticulated) in emotional speech than non-emotional speech. HAPPY tends to be produced with a more fronted vowel, and SAD, with a more backed vowel. The jaw and tongue dorsum position also tend to be lower for emotional speech, with the tongue more forward for HAPPY speech and more back (and low) for SAD speech. Duration, at least for one speaker, however, does not change as a function of the emotion, even though jaw displacement does. This suggests that jaw displacement and duration can be independent, especially during emotional speech. Numerous other interspeaker differences also are reported, and more work is needed, especially data analysis of more speakers.

This study is but a tip of the iceberg report on various observations about some of the acoustic and articulatory characteristics of Mandarin Chinese emotional speech, specifically for phrase/word final syllables on Tone 3. Future work needs to examine the changes the effect of different tones has on emotional expressions, ideally with different sets of vowels. Although this paper is based on a small set of data, it is offered as a stepping stone toward a

better understanding of the acoustic and articulatory characteristics of emotional speech expressions in Mandarin Chinese.

Acknowledgements

This work was supported by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (C) #25370444 and (C) #25240026. A special acknowledgement to Mark Tiede for help with mview and to Jianwu Dang for the use of the EMA Lab at JAIST.

References

- Boersma, P. and Weenink, D. (2015) Praat. Available [Jul. 2015] from www.praat.org
- ChangLiao, I. and Chiang, W-Y. (2004) A study of the influences of emotion on Mandarin Tones, *Unpublished Master Thesis. National Taiwan University*, Taiwan.
- Dang, J, Li, A., Erickson, D., Suemitsu, A., Akagi, M., Sakuraba, K., Minematsu, N. and Hirose, K. (2010) Comparison of emotion perception among different cultures. *Acoustical Science and Technology*, 31, 6, 394-402.
- Erickson, D., Abramson, A., Maekawa, K., Kaburagi, T. (2000) Articulatory characteristics of emotional utterances in spoken English. *Proceedings of the International Conference of Spoken Language Processing*, 2, 365-368.
- Erickson, D. (2002) Articulation of extreme formant patterns for emphasized vowels. *Phonetica*, 59, 134-149.
- Erickson, D. (2005) Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology* 26.4, 317-325.
- Erickson, D., Yoshida, K., Menezes, C., Fujino, A. Mochida, T., and Shibuya, Y. (2006) Exploratory study of some acoustic and articulatory characteristics of *sad* speech. *Phonetica*, 63, 1-25.
- Erickson, D., Shochi, T., Menezes, C., Kawahara, H., and Sakakibara, K. (2008) Some non-F0 cues to emotional speech: An experiment with morphing. *Speech Prosody 2008, Brazil, May 2008*, 677-680.
- Erickson, D., Huang, C-F., Shochi, T., Rilliard, A., Dang, J., Iwata, R., and Lu, X. (2008) Acoustic and articulatory cues for Taiwanese, Japanese and American listeners' perception of Chinese *happy* and *sad* speech. *Proc. ASJ '2008 Fall Meeting*, 1-Q-14.
- Erickson, D. Menezes, C., Sakakibara, K. (2009) Are you laughing, smiling or crying? *APSIPA, Hokkaido, October 2009*, 531-537.
- Erickson, D. and Kawahara, S. (2016) Articulatory correlates of metrical structure: Studying jaw displacement patterns. *Linguistic Vanguard* 2.
- Erickson, D., Iwata, R., and Suemitsu, A. (2016) Jaw displacement and phrasal stress in Mandarin Chinese, *TAL 2016*.
- Gu, W. and Lee, T., (2007) Quantitative Analysis of F0 Contours of Emotional Speech of Mandarin, *Proc. 6th ISCA Speech Synthesis Workshop*, 228-233.
- Kawahara, S., Erickson, D., and Suemitsu, A. (submitted) A quantitative study of jaw opening: An EMA study of Japanese vowels. *Journal of International Phonetic Association*.
- Ladefoged, P. (2005) *A Course in Phonetics*, 5th Ed. Thomson/Wadsworth Publishers.
- Li, A., Fanq, Q, Hu, F., Zheng, L. Wang, H., Dang, J. (2010) Acoustic and articulatory analysis on Mandarin Chinese vowels in emotional speech. *Proc. Institute of Electrical and Electronics Engineers*, 38-43.
- Lin, H-Y. and Fon, J. (2012) Prosodic and acoustic features of emotional speech in Taiwan Mandarin. *Proceedings of 6th International Conference on Speech Prosody*, 450-453.
- Liu, P., Pell, M.D. (2012) Recognizing vocal emotions in Mandarin Chinese: a validated database of Chinese vocal emotional stimuli. *Behav. Res. Methods*, 4, 1042-51. doi:

10.3758/s13428-012-0203-3.

- Menezes, C., and Erickson, D. (2013) Intrinsic variations in jaw deviation in English vowels. *Proc. of International Congress of Acoustics. Proceedings of Meetings on Acoustics*, **19**, 060253.
- Nguyen, B.P., Tokuda, I., and Erickson, D. (2008) Analysis of the roles of glottal features for emotion classification in spontaneous and acted emotional signals. *Proc. ASJ '2008 Fall Meeting*, 1-Q-20.
- Ohala, J. J. (1994) The frequency codes underlies the sound symbolic use of voice pitch. In L. Hinton, J. Nichols, & J. J. Ohala (eds.), *Sound symbolism*. Cambridge: Cambridge University Press. 325-347.
- Scherer, K. (1989) Vocal correlates of emotional arousal and affective disturbance. In H. Wanger and A. Manstead (eds.) *Handbook of Social Psychophysiology*, John Wiley & Sons, Ltd.
- Scherer, K. R., Banse, R., and Wallbott, H. G. (2001) Emotion inferences from vocal expression correlate across languages and cultures, *J. Cross-Cultural Psychol.*, **32**, 76-92.
- Shochi, T., Rilliard, A., Auberge, V. Erickson, D. (2009) Intercultural Perception of English, French and Japanese Social Affective Prosody. *Affective Prosody*, (ed.) M. Gotti, series: Linguistic Insights, Peter Lang Publishing Group, New York, Bern, Berlin, Bruxelles, Frankfurt am Main, Oxford, Wien, pp. 31-60.
- Spring, C., Erickson, D. and Call, T. (1992) Emotional modalities and intonation in spoken language. *Proceedings of the International Conference on Spoken Language Processing*, 679-682.
- Szameitat, D. P., Darwin, C. J., Szameitat, A. J. , Wildgruber, D. , Sterr, A., Dietrich, S., Alter, K. (2007) Formant characteristics of human laughter. *Interdisciplinary Workshop on the Phonetics of Laughter*, Aug. 4,5.
- Wang, H., Li, A., and Fang, Q. (2005) F0 contour of prosodic word in happy speech of Mandarin, *Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science*, **3784**, 433-440.
- Wen, M., Wang, M., Hirose, K., Minematsu, N. (2011) Prosody conversion for emotional Mandarin speech synthesis using the tone nucleus model. *IPSJ SIG Technical Report* vol. 2011-SLP-87 No. 2 2011/7/21.
- Williams, J.C., Erickson, D., Ozaki, Y., Suemitsu, A., Minematsu, N., Fujimura, O. (2013) Neutralizing differences in jaw displacement for English vowels, *Proc. of International Congress of Acoustics*. POMA **19**, 060268.
- Yang, Y. Wu, Z., Wu, T. Li, D. (2007) Mandarin Affective Speech. LDC2007S09, ISBN: 1-58563-442-5.
- Yuan, J., Shen, L. and Chen, F., (2002) The acoustic analysis of anger, fear, joy and sadness in Chinese, *Proc. 7th International Conference on Spoken Language Processing*, 2025-2028.
- Zhang, S., Ching, P.C., Kong, F. (2006) Acoustic analysis of emotional speech in Mandarin Chinese. *International Symposium on Chinese Spoken Language Processing (ISCSLP 2006*.