ACOUSTIC AND ARTICULATORY ANALYSIS ON JAPANESE VOWELS IN EMOTIONAL SPEECH

Mengxue CAO, Aijun LI, Qiang FANG,

Phonetics Lab, Institute of Linguistics,
Chinese Academy of Social Sciences
cmxsmile@gmail.com, {liaj; fangqiang}@cass.org.cn

Jianguo WEI, Chan SONG, Jianwu DANG

School of Computer Science,
Tianjin University
jianguo@tju.edu.cn, songchan_8855@126.com

ABSTRACT

Acoustic and Articulatory features of Japanese vowels were examined in “Neutral”, “Angry”, and “Sad” speech, using NDI Wave System. The results suggest that (1) Significant differences of the acoustic space, measured by F1 and F2, exist among different emotions. “Angry” is characterized by a horizontally compressed acoustic space, while “Sad” is characterized by a vertically compressed acoustic space. (2) The “front raising” and “retraction and back raising” patterns of the tongue movement mechanism can be enhanced by “Angry” and “Sad” emotion. (3) The lips’ dynamically protruding feature is shared by both “Angry” and “Sad”, apart from the exception [A]. We suggested that the exception is caused by the increase of the mouth opening. The mouth opening and the degree of lip protrusion are a pair of complementary features. (4) In articulatory domain, “Angry” is characterized by an increase of mouth opening and a reducing of tongue horizontal movement range.

Index Terms **—** Acoustic, articulation, Japanese vowels, emotion, NDI Wave System

1. INTRODUCTION

Emotional information has been raised as an important factor for improving the expressiveness of synthesized speech and the accuracy of natural speech recognition. Acoustic as well as prosodic features in emotional speech have been widely investigated. To better understand their underlying characteristics, with the help of modern techniques, articulatory features have become increasingly detailed investigated and closely associated with emotional information [1, 3, 4, 10-12].

Hu [5] examined the acoustic and articulatory aspects of vowel production in Ningbo Chinese, using EMA. He pointed out that complicated tongue shape data can be successfully decomposed into two underlying lingual movement mechanisms, namely the “front raising” and the “retraction and back raising” [5]. However, emotional effects on this theory remain unexplored.

Li *et al.* [1] examined the acoustic and articulatory features of Mandarin Chinese vowels in emotional speech, using EMA. Lips’ dynamically protruding feature was reported to be shared by both “Angry” and “Sad”.

In this study, emotional speech of “Neutral”, “Happy”, “Angry” and “Sad” in Japanese were collected by the NDI Wave Speech Research System [6, 8]. The emotional influences on vowel production were analyzed in both acoustic domain and articulatory domain, and the relations between acoustic and articulatory features of vowel production in emotional speech were investigated.

2. METHOD

2.1. Data and speaker

There are 5 vowels in Japanese, *a* [A], *i* [i], *u* [ɯ], *e* [e], and *o* [o]. In this study, for each vowel, apart from the vowel itself and its long-vowel form, we chose 12 monosyllable words consisted of initial consonants from 3 bilabials *p*, *b*, *m*, 6 alveolars *s*, *z*, *t*, *d*, *n*, *r*, and 3 velars *k*, *g*, *h*. As a result, for each vowel, we had 14 words (i.e. *a*, *aa*, *pa*, *ba*, *ma*, *sa*, *za*, *ta*, *da*, *na*, *ra*, *ka*, *ga*, *ha*) with little emotional preference. In total, 70 words (14 words × 5 vowels) were listed as our recording scripts.

An experienced female native Japanese speaker (42 years old) was required to produce those words in 4 different emotions, “Neutral”, “Happy”, “Angry” and “Sad”.

2.2. NDI Wave Speech Research System recording

The movements of 6 articulatory positions, Tongue Rear (TR), Tongue Blade (TB), Tongue Tip (TT), Lower Incisor (LI), Lower Lip (LL) and Upper Lip (UL), were tracked in 3D space, by the trajectory of the corresponding 5DOF sensors glued on them (Fig. 1). Another 3 5DOF sensors were glued to the bridge of nose (NOSE), the back of left ear (LE), and the back of right ear (RE) respectively as reference points. Those points were relatively stable during the recording, so they were used as reference to remove the head movements of the speaker.

978-1-4673-2507-3/12/$31.00 © 2012 IEEE

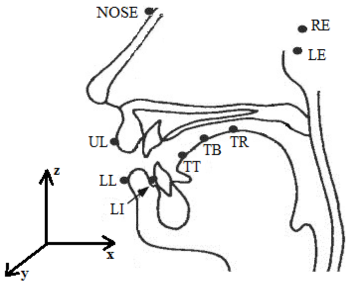


Fig. 1. Sensor position and the coordinate system

For each emotion, the speaker read all 70 words in a row with short pauses in between. The acoustic data and the articulatory data were collected simultaneously by the NDI Wave System. For 3D tracking, it had a sampling frequency of 100 Hz, a static positional accuracy of 0.6mm, and a dynamic positional accuracy of 1.5mm. The sampling frequency for the acoustic signal was 22 kHz.

2.3. Data post-processing

2.3.1. Acoustic data

The acoustic data was saved as WAV files. Each consonant and vowel was labeled and chopped using Wavesurfer [7], and only vowel parts were kept for experimental analysis. The values of the first and second formant (F1 and F2) of each vowel clip were automatically extracted from the audio data using Praat [9]. Then, for each of the 5 vowels, the average F1 and F2 value under each emotion were calculated respectively.

2.3.2. Articulatory data

Based on the fact that the dynamic positional accuracy of the NDI Wave System we used was 1.5 mm, a rounding operation was first performed throughout the articulatory data. Then, based on the fact that the sampling frequency of the System was 100 Hz, a 5 Hz third-order Butterworth low-pass filter was applied to data of the 3 reference points, while a 20 Hz third-order Butterworth low-pass filter was applied to data of the 6 articulatory points.

To remove the influence of head movements of the speaker during recording, a head-correction operation was applied to data of the 6 articulatory points based on the reference data from the 3 reference points, using MatLab scripts.

Then, a coordinate shifting of all data points was done by taking the NOSE reference point as the origin of the coordinate system. Only in this way could the coordinate-data-based comparison between different emotions be meaningful.

For each articulatory tracking point, only the front-back horizontal movements (i.e. *TRx*) and the high-low vertical movements (i.e. *TRz*) were kept; for the movement trajectory of each articulatory point, only the middle part, which was stable, was kept. The coordinate mean value of each trajectory was used for experimental analysis.

3. PERCEPTION TEST

The recording of each emotion was re-chopped every 5 words (i.e. *a i u e o/ ka ki ku ke ko/ pa pi pu pe po*), so that enough emotional information could be included in the utterance presented to the subjects.

8 native Japanese speakers (4 males and 4 females, average age 22.13, max. 25, min. 20) were participated as subjects. Each subject was required to listen to all the 56 utterances (14 utterances × 4 emotions) and indentify the intended emotion of the utterances.

The results are shown in Table 1. The intended emotion "Neutral", "Angry" and "Sad" are well perceived, while the perceiving of "Happy" is not good enough. 22% of the intended "Happy" utterances are perceived as "Neutral", and 18% are as "Angry".

Xu has summarized such kind of misjudgment between happy and angry in detail [11]. Although happy and angry are among the most distinguishable emotion pairs in daily communications, they are among the least distinguishable pairs in terms of acoustic features. Xu argued that the size-code is the critical cue to distinguishing "Happy" and "Angry" [11, 12]. However, our perception experiment shows that a same degree of misjudgment also exists between "Happy" and "Neutral". We can hardly put a conclusion on whether those phenomena were caused by reasons of perception domain, or by the less expressiveness of the recording itself. In this case, we decide to drop out the "Happy" emotion in the analysis sections, and we would like to deep explore those phenomena in our further studies.

Table 1. The confusion matrix of 4 emotions. "Int." stands for intended emotion, "Per." stands for perceived emotion. All values have been normalized into percentage representation. The perceiving of "Happy" is not good enough.

Per. Int.	Neutral	Happy	Angry	Sad
Neutral	**0.88**	0.03	0.04	0.05
Happy	0.22	0.58	0.18	0.02
Angry	0.09	0.04	**0.86**	0.01
Sad	0.01	0.00	0.00	**0.99**

4. RESULTS

4.1. Acoustic analysis

The average values of F1 and F2 for vowels in emotion "Neutral", "Angry" and "Sad" are shown in Fig. 2 in Bark scale. Significant differences can be found in the acoustic space across different emotions.

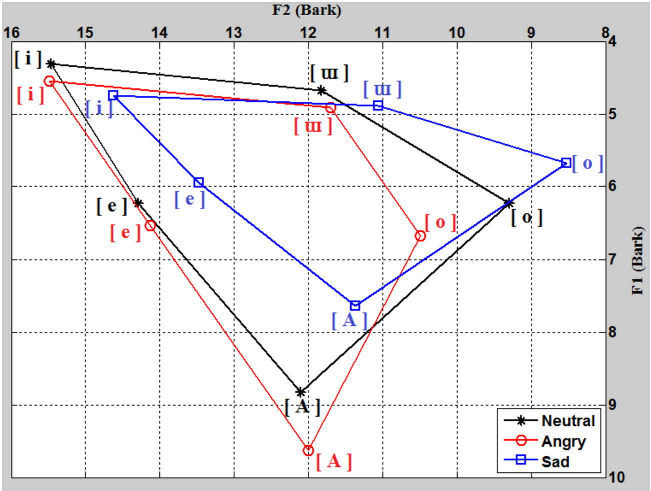


Fig. 2. Acoustic space plots for vowel [A], [i], [ɯ], [e], [o] in emotion "Neutral", "Angry" and "Sad". The vertical coordinate is F1, and horizontal coordinate is F2. Bark is the measure scale.

In "Angry" emotion, compared with "Neutral", the acoustic space moves downwards, which indicates a bigger mouth opening. Horizontally, the mid-back vowel [o] moves significantly forward, and the mid-front vowel [e] moves a little backward. In other words, the acoustic space has been compressed in horizontal dimension, implying a horizontally tighter tongue movement.

In "Sad" emotion, compared with "Neutral", the acoustic space moves significantly backwards, which indicates a strong tongue retraction. Vertically, the high vowel [i] and [ɯ] move downwards, while the mid vowel [e] and [o], and the low vowel [A] move upwards. In other words, the acoustic space has been compressed in vertical dimension, implying a vertically tighter tongue movement.

4.2. Articulatory analysis

The front-back and high-low relations of the articulatory data for each vowel in 3 different emotions were investigated. Since the tongue contours of [A] and [i] are of the most representative among the five vowels, we took the plots of the two as examples to be shown in Fig. 3.

To better investigate the articulatory feature, we defined 4 parameters based on information reflected from the data. We denoted the Mouth Opening (MO) by calculating the vertical difference between sensor UL and LL:

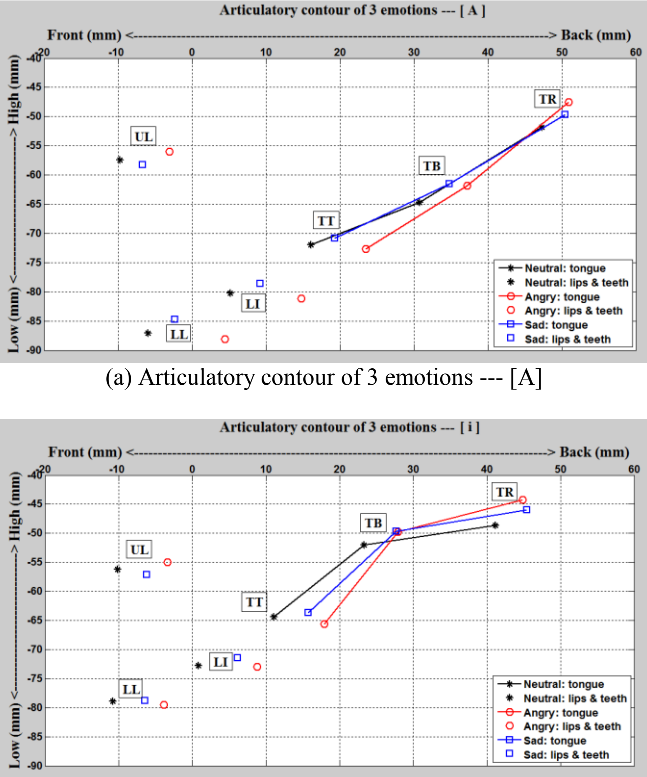
$$MO = ULz - LLz \qquad (1)$$

We denoted the degree of Lip Protrusion (LP) by calculating the horizontal difference between sensor LI and LL:

$$LP = LIx - LLx \qquad (2)$$

We denoted the length of the Front Oral Cavity (FOC) by calculating the horizontal difference between sensor TR and LL:

$$FOC = TRx - LLx \qquad (3)$$



(a) Articulatory contour of 3 emotions --- [A]

(b) Articulatory contour of 3 emotions --- [i]

Fig. 3. Articulatory space plots of emotion "Neutral", "Angry" and "Sad" for vowel [A] and [i]. The vertical coordinate represents the high-low position of the tongue, and the horizontal coordinate represents the front-back position of the tongue. Millimeter (mm) is the measure scale.

We denoted the Range of the Tongue Movements in Horizontal Dimension (RoH) by calculating the horizontal difference between sensor TR and LI:

$$RoH = TRx - LIx \qquad (4)$$

Every parameter was then normalized by dividing the maximum value of that parameter. For example:

$$MO_{normalized:[A]\sim Sad} = \frac{MO_{[A]\sim Sad}}{\arg\max(MO_{j\sim k})} \qquad (5)$$

$$where, \quad j=\{[A], [i], [ɯ], [e], [o]\},$$
$$k=\{Neutral, Angry, Sad\}$$

Compared with "Neutral", we find clear tendencies for all 5 vowels that all articulators move upwards vertically and backwards horizontally, in emotion "Angry" and "Sad".

From Fig. 4, we can see that for all 5 vowels, the range of the horizontal tongue movements (RoH) decreases from "Neutral" to "Sad" and to "Angry". Moreover, "Angry" has the biggest mouth opening (MO), while "Neutral" secondary and "Sad" third. By comparing the MO and LP results, we can find a complementary relation between them. The increase of mouth opening causes the decrease of lip protrusion, while the decrease of mouth opening leads to the increase of lip protrusion.

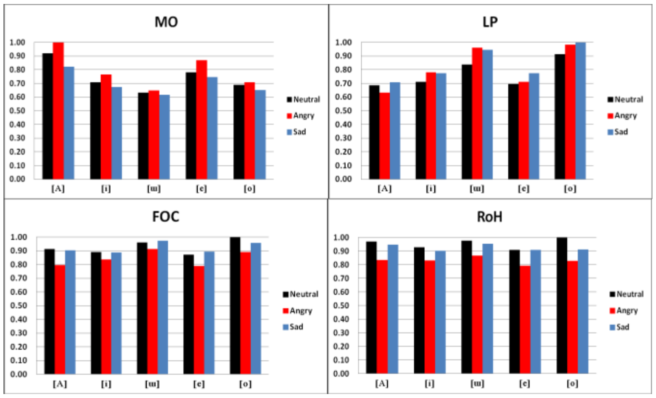


Fig. 4. The percentage representation of the parameters MO, LP, FOC and RoH, among emotion "Neutral", "Angry" and "Sad" for vowel [A], [i], [ɯ], [e], [o]. Each parameter has been normalized.

For low-central vowel [A], compared with "Neutral", both the TB and TR sensor move backwards and upwards, and TT moves backwards, in emotion "Angry" and "Sad". As shown in Fig. 4, the mouth opening is the biggest in "Angry", and much smaller in "Sad" with 18% drop, while the lips protrude most in "Sad", and least in "Angry".

For high-front vowel [i], in emotion "Angry" and "Sad", both the TB and TR sensor move backwards and upwards, and TT moves backwards. Those indicate that the tongue body rises higher towards the palate. Fig. 4 shows that the lips protrude more in "Angry" and "Sad".

For high-back unrounded vowel [ɯ], the tongue contour is very similar to that of [A]. The major difference between them is that the vertical tongue position of [ɯ] is about 5mm higher than that of [A]. Lips protrude significantly in "Angry" and "Sad", and the mouth opening stays consistent across 3 emotions.

For mid-front vowel [e], the tongue contour is very similar to that of [i]. The major difference is that the TB of "Angry" is lower than that of "Sad". The mouth opens bigger in "Angry", while the lips protrude more in "Sad".

For mid-back vowel [o], TT moves backwards, TB moves backwards and upwards, and TR moves upwards, in emotion "Angry" and "Sad". Compared with "Neutral", the range of the tongue horizontal movements reduces significantly in "Angry" with 17% drop, and moderately in "Sad" with 9% drop.

4.3. Comparative analysis

In "Angry" emotion, the downward movement of the acoustic space is consistent with its biggest mouth opening among 3 emotions. The horizontally compressed acoustic space is consistent with the shortest range of the horizontal tongue movement.

In "Sad" emotion, the vertically compressed acoustic space is consistent with its smallest mouth opening among 3 emotions. The backward movement of the acoustic space can be seen as the co-effect of the length of the front oral cavity and the degree of lip protrusion.

5. DISCUSSIONS AND CONCLUSIONS

In the perception test, we find that the speech with intended "Happy" emotion can be misjudged into both "Neutral" and "Angry". Two main questions are raised, (1) What is the underlying reason for the misjudgment of "Happy"? (2) Compared with findings in [11], why "Happy" is misjudged as "Neutral" and "Angry" in a same degree, and why "Neutral" and "Angry" are seldom misjudged as "Happy"?

In acoustic domain, our findings are consistent with Li [1] and Breazeal [2]. "Angry" causes strong high-frequency energy, and "Sad" leads to little high-frequency energy. Comparing the shape of the acoustic space, we find that the "compress direction" is a distinctive feature between "Angry" and "Sad". "Angry" is characterized by a horizontally compressed acoustic space, while "Sad" is characterized by a vertically compressed acoustic space.

In articulatory domain, we have found similar patterns of "front raising" and "retraction and back raising" of the tongue movement mechanism, stated by Hu [5]. We also find that those patterns can be enhanced by emotion "Angry" and "Sad". In "Angry" and "Sad" speech, for front vowel [i] and [e], the front part of the tongue raises higher, which increases the "front raising" effect; for central vowel [A], and back vowel [ɯ] and [o], the tongue retracts more and the back part of the tongue rises higher.

Xu argued that the vowels synthesized with dynamically protruding lips are more frequently heard as angry or spoken by a larger person [12]. Li claimed that the same hypothesis also applies to "Sad" [1]. Our findings suggest similar conclusions to Li's, apart from the exception [A]. As shown in Fig. 4, lips protrude least in "Angry" speech of [A]. Taking the mouth opening into consideration, we suggest that the decrease of the lip protrusion is caused by the increase of the mouth opening, while a speaker is trying to express a high arousal emotion on an open vowel. The mouth opening and the degree of lip protrusion are a pair of complementary features.

As reported by Li [1], Lee [10] and Erickson [3], we find that "Angry" can be characterized by big mouth opening. Moreover, our articulatory data suggest that the reducing of horizontal tongue movement range is also an important feature of "Angry".

6. ACKNOWLEDGEMENTS

This research was funded by the JSPS Ronpaku Program, the NSFC Project with No. 60975081, and the CASS Innovation Program.

7. REFERENCES

[1] A. Li, Q. Fang, F. Hu, L. Zheng, H. Wang, and J. Dang, “Acoustic and Articulatory Analysis on Mandarin Chinese Vowels in Emotional Speech”, *ISCSLP2010*, Tainan, 2010.

[2] C. Breazeal, *Designing Social Robots*. MIT Press, Cambridge, 2001.

[3] D. Erickson, A. Abramson, K. Maekawa, and T. Kaburagi, “Articulatory Characteristics of Emotional Utterances in Spoken English”, *6th ICSLP*, Vol. 2, pp.365-368, Beijing, 2000.

[4] D. Erickson, K. Yoshida, C. Menezes, A. Fujino, T. Mochida, and Y. Shibuya, “Exploratory Study of Some Acoustic and Articulatory Characteristics of Sad Speech”, *Phonetica*, 63(1):1-25, 2006.

[5] F. Hu, “An Acoustic and Articulatory Analysis of Vowels in Ningbo Chinese”, *Eurospeech*, 2003.

[6] K. Rudy, “The Effect of Palate Morphology on Onsonant Articulation in Healthy Speakers”, *MSc thesis of the University of Toronto*, Canada, 2011.

[7] K. Sjölander and J. Beskow, *Wavesurfer*, School of Computer Science and Communication of KTH Royal Institute of Technology, Sweden, retrieved in 2011.
http://www.speech.kth.se/wavesurfer/

[8] Northern Digital Inc., *Wave User Guide*, Northern Digital Inc., Canada, 2011.

[9] P. Boersma and D. Weenink, *Praat*, Phonetic Sciences of University of Amsterdam, Netherland, retrieved in 2012.
http://www.fon.hum.uva.nl/praat/

[10] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, “An Articulatory study of Emotional Speech Production”, *Eurospeech*, 2005.

[11] Y. Xu, A. Kelly and C. Smillie, (in press), “Emotional expressions as communicative signals”, *To appear in S. Hancil and D. Hirst (eds.) Prosody and Iconicity*, 2012

[12] Y. Xu and S. Chuenwattanapranithi. “Perceiving Anger and Joy in Speech through the Size Code”. In *Proc. 16th ICPhS*, Saarbrucken, pp. 2105-2108, 2007.