Investigation of the relation between acoustic features and articulation — an application to emotional speech analysis

Yongxin WANG*†, Jianwu DANG†‡ and, Lianhong CAI*
*State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology(TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing
†School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan
‡School of Computer Science and Technology, Tianjin University, Tianjin

***Abstract*—In speech communication, humans are able to perceive not only the acoustic features but also the articulation state from speech sounds. Some specific articulation configurations can carry certain non-linguistic information such as emotion. However, the relation between articulation and acoustic features is still not clear. To clarify such a relation, we analyzed speech data from the emotions of cold anger and hot anger based on an emotional speech corpus and attempt to realize the observations using a physiological articulatory model. We measured the changes in acoustic features when altering articulation states for individual speech organs using the physiological articulatory model. It is found that the cold and hot angers both have emphasized components in high-frequency region, while the former controls articulation and the latter controls the sound source of the speech production system.**

I. Introduction

Most of important non-linguistic information in speech communication, such as emotion, intention, etc., is concerned with some specific articulation state. Since the relation between articulation and acoustic features is not clear, it has not been considered sufficiently yet in speech analysis, though humans are able to recognize articulation states as well as acoustic features. In this study, the authors attempt to clarify such a relation and apply it to the analysis of emotional speech.

Human speech production system can be approximated as a source-filter system, where the glottal source is treated as the source for voiced sounds, and the time-varying vocal tract is treated as the filter. Controlling of vocal tract, the filter of the speech production system, would generate some special features in speech sound. Although listeners are able to spot this kind of acoustic features easily, and figure out how they are produced by articulation, this kind of features have not been well studied yet. Clarifying the relation between articulation and acoustic features possibly provide knowledge to extract and manipulate the acoustic features in speech synthesis and speech recognition, as well as understanding the underlying mechanisms of human speech production.

In communication, spoken language carries more information than written language, since plenty of non-linguistic information is added by controlling the vocal system in a specific way. For example, the speaker would manipulate the vocal system in different ways consciously or unconsciously based on his/her emotional state. The changes in sound source and vocal tract configuration add some special features to speech sound. Those changes are cues for listeners to perceive the emotion state of the speaker.

In this work, an emotional speech corpus is first analyzed to obtain some evidence of emotion related acoustic features. An analysis-by-synthesis method is employed in this investigation. With the possible articulatory states that would be used in emotional states, a physiological articulatory model is used to approach the acoustic features found in real speech by manipulating the articulators individually [1], [2]. The relation of articulation and acoustics is revealed using the articulatory and acoustic data in the simulation data set based on a model based analysis-by-synthesis method.

II. Analysis of Emotional Speech

To obtain certain evidence for the modal based study, we first analyzed an emotional speech corpus recorded in Fujitsu Lab to extract emotion-related acoustic features. The corpus contains five emotions, including neutral, cold anger, happy, sad, and hot anger. Twenty Japanese sentences were performed by one female speaker, and recorded with a sample rate of 22050 Hz. In this study, we mainly focus on the speech materials with the emotions of hot and cold angers, while the neutral one is used for comparison.

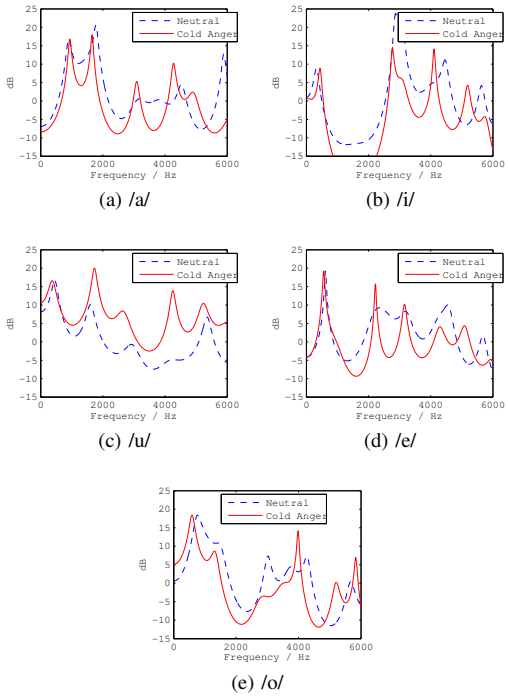
The sentences with those three emotions are labeled in phoneme level, all the vowels were extracted from the sentences for analysis. For each emotion, the numbers of five Japanese vowels are listed in Table I. In analysis, the speech signals are first resampled to 16 kHz. After pre-emphasis, twenty-order LPC coefficients are calculated for each frame with a 25-ms Hamming window and a 12.5-ms shift. The spectrum envelopes is calculated from the LPC coefficients for each vowel [3].

The typical spectra of neutral and cold anger speech sounds are shown in Figure 1 for five Japanese vowels. Since the volume level for each emotion is different in the recording, the

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

TABLE I
THE NUMBER OF APPEARANCES OF EACH VOWEL IN THE CORPUS, AND THE PERCENTAGE AMONG ALL VOWELS

Vowel	/a/	/i/	/u/	/e/	/o/
No.	91	55	31	26	34
%	38.4%	23.2%	13.1%	11.0%	14.3%



(a) /a/ (b) /i/ (c) /u/ (d) /e/ (e) /o/

Fig. 1. Typical spectrum envelope shape for vowels in cold anger

intensity levels are normalized by setting the first formant to the same amplitude when comparing the spectrum envelopes.

It is found that the vowels with the emotion of cold anger show a distinctive structure in high frequency region comparing to the neutral one. The speech sounds with cold anger have sharper resonance peaks than neutral sounds in high frequency region. Since in neutral speech the high frequency region is deteriorated because of less power in the voiced sound source or irregularity of the harmonics, sharpening resonance peaks in high frequency region is of benefit to emphasizing the component in that region. That is, it seems that cold anger emphasizes the component of this frequency region by sharpening the resonance peaks.

For further investigation, we averaged the bandwidths of the formants over the region between 3 kHz and 5 kHz for each vowel, which were estimated using the complex root pairs of the LPC polynomial. Comparing the bandwidths between cold anger and neutral speech, it is found that for /a/ and /e/ the formants of cold anger have narrower bandwidths than those in neutral speech, with a statistical significance level of 0.05. The bandwidths are similar for other vowels between cold anger and neutral speech, in a statistic point of view. This shows in cold anger, the open vowels gets an emphasized high

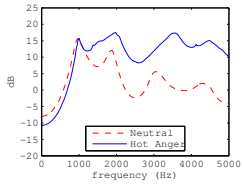


Fig. 2. Average spectrum envelopes with normalized energy from hot anger and neutral speech (/a/)

frequency region with narrower bandwidths of the formants in that region.

The same analysis approach was also applied to the speech sound with hot anger. An average spectrum of /a/ is shown in Figure 2. Comparing with neutral sound, hot anger speech sound possesses more energy in high frequency region. This can be seen for most vowels in hot anger speech. This phenomenon is consistent with the report of previous works [4].

III. PHYSIOLOGICAL ARTICULATORY MODEL

Based on the above observation, we investigated the relation between acoustic feature and articulatory feature using an analysis-by-synthesis method with a physiological articulatory model. The physiological articulatory model consists of the tongue, jaw, vocal tract wall, and lips, which are driven by target-based muscle contraction, and the synthesis part uses the transmission line model [1], [2], [5]. This model is capable of generating speech sounds for phonemes, syllables and short sentences.

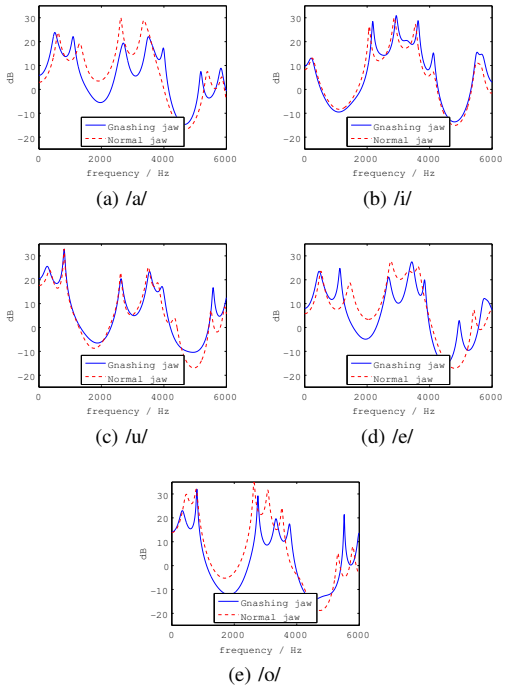
The geometry data of the articulators in the model were obtained from a male subject. The articulators are connected together by muscles, soft tissues and joints. Muscular structure was replicated according to anatomical data of human. Muscle activation patterns are estimated from the given target of the articulators and are used to drive the model. Using this model with the control strategy, we are able to simulate any possible articulation for speech production.

To synthesize speech sound, cross-sectional area function of the vocal tract is calculated from vocal tract shapes. After applying the transmission line model on the area function with a sound source, the speech sound would be produced for various articulation configurations [5]. The model of glottal area function is used to generate a voiced sound source, fricatives and stops are automatically generated based on aerodynamics at the constriction or closure in the vocal tract.

In this study, this model is controlled to simulate normal and various specific speech sounds. Different phonemes are realized by setting target positions for individual articulators according to the observation of microbeam data [6]. Emotional articulation is realized by manipulating the positions of specific articulators. The relation of acoustics and articulation is investigated using both acoustic and articulatory data from the simulation data set.

TABLE II
NORMAL POSITIONS OF THE JAW FOR FIVE JAPANESE VOWELS. REPRESENTED BY THE GAP BETWEEN UPPER AND LOWER TEETH

Vowel	/a/	/i/	/u/	/e/	/o/
gap (cm)	1.55	0.62	0.78	1.24	1.32



(a) /a/ (b) /i/ (c) /u/ (d) /e/ (e) /o/

Fig. 3. The spectrum envelopes for the vowels with the jaw in different positions

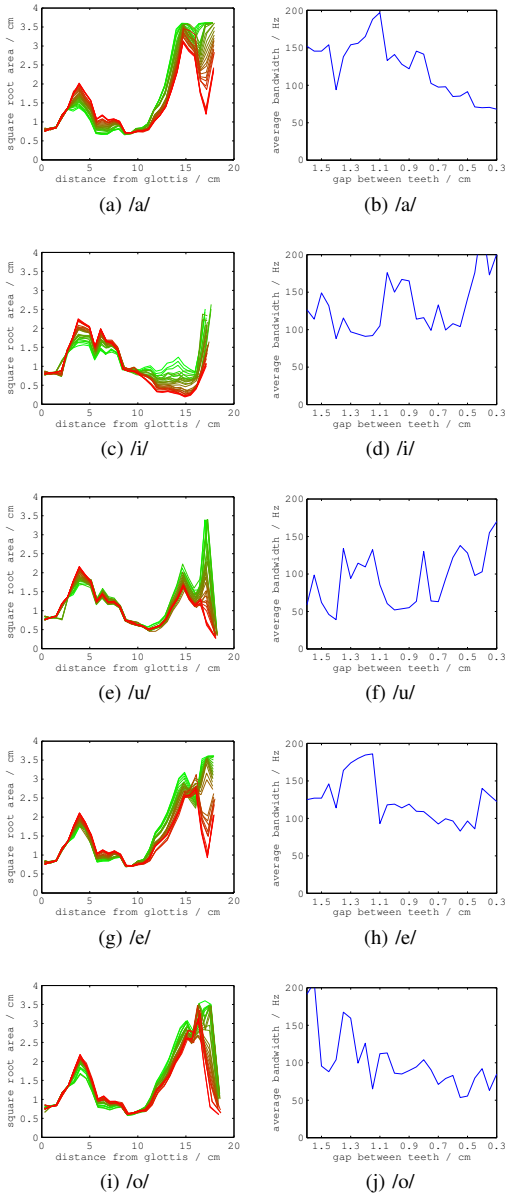
IV. SIMULATION EXPERIMENT

In the model-based numerical experiment, Vowel-Consonant-Vowel (VCV) sequences are synthesized based on appropriate initial articulation target with a certain variation.

In producing the cold anger speech, speaker usually gnashes the jaw to form a special articulation state. For this reason, the position of the jaw is a crucial factor for simulating the cold anger. The jaw position is set by the gap between the upper and lower teeth, ranged from 0.3 cm (gnashing position) to 1.6 cm (the maximum value for normal position) in the experiment. The normal positions used in the simulation experiment for the Japanese vowels are listed in Table II. Comparison of the spectrum envelopes is shown in Figure 3, which was calculated in the cases of gnashing positions and normal positions for five vowels.

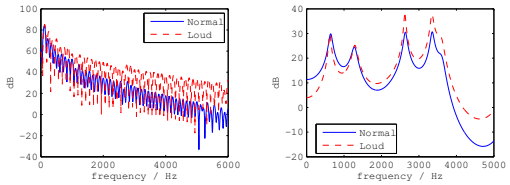
From Figure 3, one can see that the spectral structure in high frequency region shows more refinement when the jaw is in a gnashing position. This result is consistent with the observation in real data. This implies that the emphasis in the high frequency region of cold anger is caused by gnashing the jaw. This finding can also be supported by our daily experience that people usually keep the jaw in a gnashing position when producing cold anger.

The relation between jaw position, vocal tract configuration and acoustic features is further investigated. Figure 4 shows the changes in the area functions of the vocal tract when raising jaw position (reducing the gap between the teeth) in the left panel, and the average bandwidth of formants between 3 kHz and 5 kHz versus jaw position in the right panel. Since we use 30 sections to represent the cross-sectional area function of the vocal tract, some small perturbations in the tongue tip may cause discontinuous changes in the area function. One can see that the jaw position systematically affects the whole area function. In general, the area of the anterior part of the vocal tract gets smaller as the jaw is getting to a gnashing position.



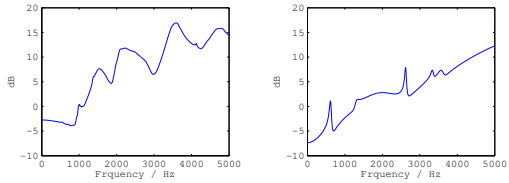
(a) /a/ (b) /a/ (c) /i/ (d) /i/ (e) /u/ (f) /u/ (g) /e/ (h) /e/ (i) /o/ (j) /o/

Fig. 4. The area functions (left) and average bandwidth in high frequency region versus the jaw position (right) for /a/ /i/ /u/ /e/ /o/, respectively. The color goes from green to red as the jaw goes up in the area function figures.



(a) (b)

Fig. 5. spectrum for the sound source (left) and spectrum envelope with normalized energy for /a/ (right) with different vocal effort



(a) (b)

Fig. 6. Difference of the (average) spectrum of /a/ between hot anger and neutral (left), and between synthesized loud and normal sound (right)

The result in the right panel shows the average bandwidth of resonance peaks between 3 kHz and 5 kHz versus jaw position. For /a/, /e/ and /o/, the average bandwidths get narrower in general as the anterior part of the area functions gets smaller by gnashing the jaw. The open vowels /a/, /e/ and /o/ do not have a narrow constriction in the anterior part of the vocal tract, and the gnashing of the jaw generates an additional constriction in the anterior part of the vocal tract. The descending tendency of bandwidth with the increase of the gnashing degree demonstrates that this additional constriction sharpens the resonance peaks in the high frequency region. In contrast, closed vowels /i/ and /u/ have narrow constrictions in the anterior part of the vocal tract. In this case, there is little effect on the acoustic property when one more constriction is added in the vicinity of the original constriction. Therefore, there is not a consistent change in the bandwidth of high frequency formants for closed vowels when the jaw position changes alone.

In hot anger speech, quite strong energy is seen in the high frequency region. This phenomenon is generally considered to be caused by changes in the property of the sound source by using more vocal effort [4], [7]. In the physiological articulatory model, the property of the sound source is controlled by the glottal area function using Fant's Model [5], [8]. Different vocal effort can be simulated by the model based on the observations in [9]. The spectrum of the sound source and the spectrum of the generated sound is shown in Figure 5 for comparison.

To get a quantitative analysis how the energy increases with the frequency, the differences of the spectrum between hot anger state and neutral state, and between the simulated loud and normal states are calculated and shown in Figure 6. One can see from the figure that with more vocal effort, the increase of energy in high frequency region in hot anger state can be approximated.

V. Conclusion

In this paper, the relation between articulatory and acoustic features of speech from cold and hot anger emotional state is studied. Emotional speech data for cold and hot angers from an emotional speech corpus were first analyzed to get emotion-dependent acoustic features. The results showed that in anger states, speakers are going to deliver every word clearly to the listener by emphasizing speech sound in high frequency region. By means of the analysis-by-synthesis method using the physiological articulatory model, we clarified that cold anger and hot anger use different control strategies to emphasize speech: for hot anger the emphasis is realized by increasing the amplitude of the sound source, while for cold anger the emphasis is realized by sharpening the resonance peaks of the high frequency region. When implementing these strategies in the model, consistent acoustic features were obtained in model simulation with those in the real speech.

In this preliminary experiment, we investigated the relation between acoustic feature and its causes in articulation in cold and hot angers. The investigation helped us further understand about how the anger speeches are generated, and the acoustic features related to them.

Acknowledgement

This study is supported in part by "Qian Ren Ji Hua" of Tianjin, China, and in part by Grant-in-Aid for Scientific Research of Japan (No. 22500150). It is also supported by National Natural Science Foundation of China (90920302, 60910130, 60928005), the National High Technology Research and Development Program ("863"Program) of China (2009AA011905).

References

[1] J. Dang and K. Honda, "A physiological articulatory model for simulating speech production process," *Acoustical Science and Technology*, vol. 22, no. 6, pp. 415–425, 2001.

[2] ——, "Construction and control of a physiological articulatory model," *The Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 853–870, 2004.

[3] J. Markel, "Digital inverse filtering-a new tool for formant trajectory estimation," *Audio and Electroacoustics, IEEE Transactions on*, vol. 20, no. 2, pp. 129–137, 1972.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[5] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, vol. 1, no. 3-4, pp. 199–229, Dec. 1982.

[6] M. Hashi, J. R. Westbury, and K. Honda, "Vowel posture normalization," *The Journal of the Acoustical Society of America*, vol. 104, no. 4, pp. 2426–2437, Oct. 1998.

[7] I. Hiradate and M. Akagi, "Analyses of acoustic features of "anger" emotional speech," *IEICE technical report. Speech*, vol. 101, no. 744, pp. 43–50, 2002.

[8] G. Fant, "Glottal source and excitation analysis," *Speech Trans. Lab. - Quarter Progress Status Report*, vol. 20, no. 1, pp. 85–107, 1979.

[9] E. B. Holmberg, R. E. Hillman, and J. S. Perkell, "Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice," *The Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 511–529, 1988.