2010 International Workshop on Nonlinear Circuits,
Communication and Signal Processing
NCSP'10, Waikiki, Hawaii, March 3-5, 2010

Investigation of Acoustic Features for Emotional Speech using a Physiological Articulatory Model

Yonxin Wang[1,2], Jianwu Dang[1,3] and Lianhong Cai[2]

1. Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa
923-1292, Japan
Phone: +81-761-51-1235
Email: {s0922001, jdang}@jaist.ac.jp

2. Tsinghua University
Tsinghua University, Haidian
District, Beijing 100084, China
Phone: +86-10-62771587
Email:
wangyongxin@mails.tsinghua.edu.cn,
clh-dcs@tsinghua.edu.cn

3. Tianjin University
92 Weijin Road, Nankai District,
Tianjin 300072, China
Phone: +86-22-27406149
Email: dangjianwu@tju.edu.cn

Abstract

In emotional speech studies, it is well known that loudness, pitch, position and length of pauses, etc. are important factors in expressing emotions. Besides those phonation features, the articulation of the vocal tract endows some spectral features that reflect emotions. Those features come from the special way of controlling articulatory organs in emotional states. Daily experiences tell us that some specific control strategies in articulation can result in certain emotional states. However, the relation between the articulation and acoustic features is not clear yet. In this study, we investigate the spectral features using a physiological articulation model by manipulating the configurations of speech organs and synthesizing the corresponding speech sounds, in an analysis-by-synthesis way. The acoustic features concerned with emotional states are clarified by comparing the simulation results with an emotional speech corpus. The results show that the changes of spectrum envelop in the simulation results are consistent with those in real speech sound.

1. Physiological Articulatory Model

A physiological articulatory model is a computational model for human articulatory organs, which can simulate the physical movements of the speech articulatory organs, and synthesize speech sound accordingly. The model used in this paper includes the articulatory organs from the larynx to the lips and nasal cavity. The control mechanism used in the model uses target positions of articulator to estimate muscle stimulations, and the articulators are then driven by muscle stimulations to the target position. We used this model to simulate various articulator configurations that are related to emotions.[1, 2]

The synthesis part of the model uses a source-filter model which is commonly used in speech research. Characteristics of the source (glottal air flow for vowels) and the filter (configuration for the vocal tract) are specified or simulated for each phoneme, then the transmission line model[4] is used to generate speech sound.

With this physiological articulatory model, phonemes, continuous phoneme sequences and sentences can be synthesized using appropriate parameters. We can then control the articulatory organs manually to realize a configuration of the vocal tract and to clarify how the acoustic output is affected by different articulatory organ configurations[3, 6].

2. Analysis of the emotional speech database

In emotional state, people would have a special way of controlling the body, including the articulatory organs. This would affect the speech sounds generated in emotional states, and cause the speech to be heard as emotional[5]. For example, in cold anger, an emotional state that an angry person tries not to outburst, the articulatory organs may not move in great amount. That would result in that the jaw does not always open large enough for an open vowel. In hot anger, people would go into an outburst and the effort used causes the sub-glottal pressure to increase. In both situations, people are able to hear the angry emotion in the speech, which means there may be some common acoustics features reside in these two different angry emotions.

To clarify such features we first analyzed an emotional speech corpus recorded by Fujitsu Laboratory. The corpus contains speeches read in five different emotions, which are neutral, happy, cold anger, sad and hot anger, by one female speaker. Twenty Japanese sentences are performed in the five emotions. The recording for each sentence is then labeled, and the spectrum contour is calculated for each phoneme using LPC. The vowels and semi-vowels are analyzed in this study as they are affected the most by the control of glottal air flow and the shape of the oral cavity.

The average normalized spectrum envelop of all vowels is

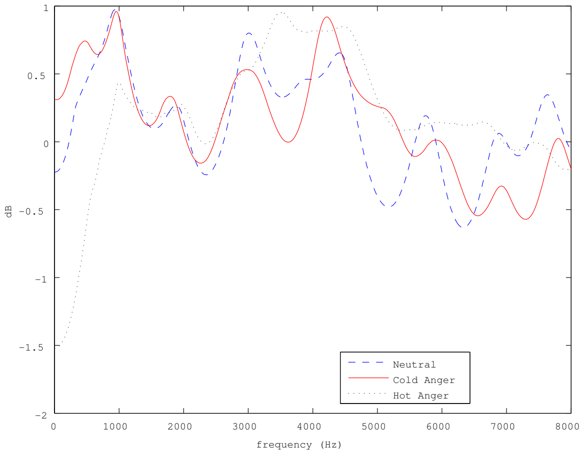


Figure 1: Average normalized spectrum envelops of cold and hot angers compared with the neutral one

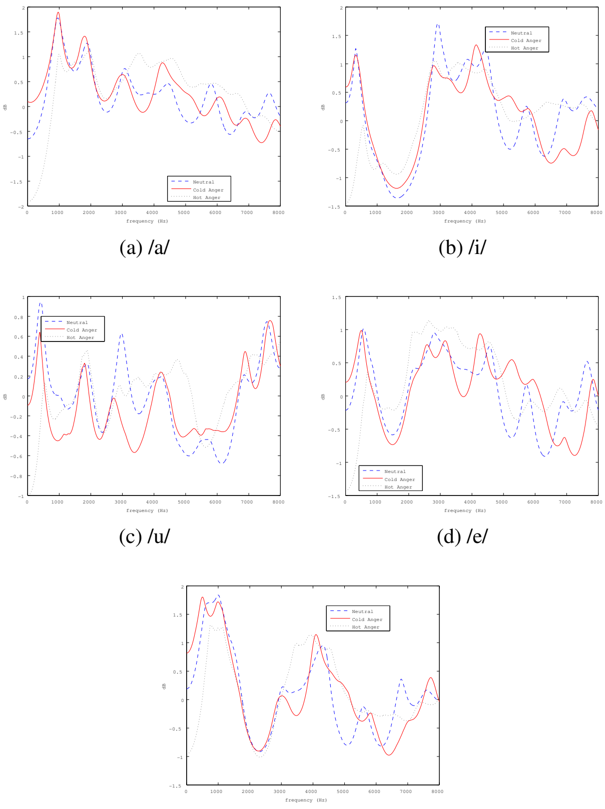
shown in Figure 1. The integrity is normalized for comparison. It is obvious to see that in hot anger emotion state, the whole spectrum slope has changed. This is considered to be a result of the increased effort used for speaking, which causes the waveform of the glottal air flow to become sharper. The sharper the glottal air wave the more energy in high frequency region.

In the cold anger emotional state, one can see that there is a stronger peak at around 4 kHz. It is not as significant as in the hot anger emotional state. In hot anger state, the extra power in the high frequency region can be seen for all syllables of the sentence, while in cold anger state, the similar phenomena are only seen for some syllables playing important roles in emotion expression with the special configuration. Also, the cold anger emotional state affects different vowels differently. Figure 2 shows the average spectrum envelope of all the appearances of each of the five vowels of Japanese in the emotional speech corpus. It can be seen for /a/, the extra energy increase around 4 kHz would be more obvious, but it is not the case for /o/ and /u/.

From the above results, one can see a consistent tendency for two types of anger states that the power increases in the high frequency region. What is the difference in speech production point of view? Intuitively, we speculate that in hot anger, the increased power is caused by the changes in the sound source, while in cold anger, the increased power is due to special articulation of the vocal tract.

3. Simulation Experiment

To clarify our speculation on the causes of these two kinds of emotions, we conducted a numerical experiment. In the simulation experiment, the configurations of the articulation and the way of phonation are controlled for the physiological articulatory model to generate specific sounds, and then the acoustic features are analyzed to investigate the relationship between the articulatory features related to emotions and the acoustic output.



(a) /a/

(b) /i/

(c) /u/

(d) /e/

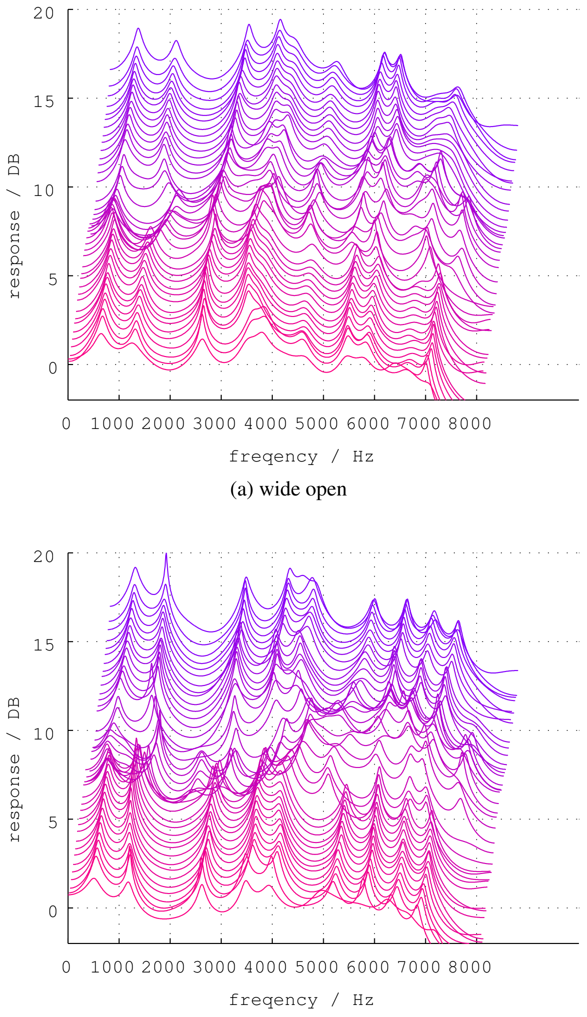
(e) /o/

Figure 2: Average normalized spectrum envelops for the five Japanese vowels

To simulate the observations in the emotional speech corpus, we controlled the factors according to our speculation to generate the anger speeches and the neutral speech. For cold anger, the relation between jaw position and the output spectrum is analyzed by setting the jaw to several different positions, from wide open to nearly closed, including the normal ones. For hot anger, the sub-glottal air pressure is set to different levels. The two experiments are carried out separately and the CVC phoneme sequences /aga/, /igi/, /ugu/, /ege/ and /ogo/ are used in both synthesis experiments.

For analysis, the glottal air wave, frequency consonance of the vocal tract and the spectrum envelop of the output sound are investigated for each simulation.

The running spectra for the phoneme sequence /aga/ are shown in Figure 3. An extra formant can be seen around 4 kHz when the jaw is in a nearly closed position. In this

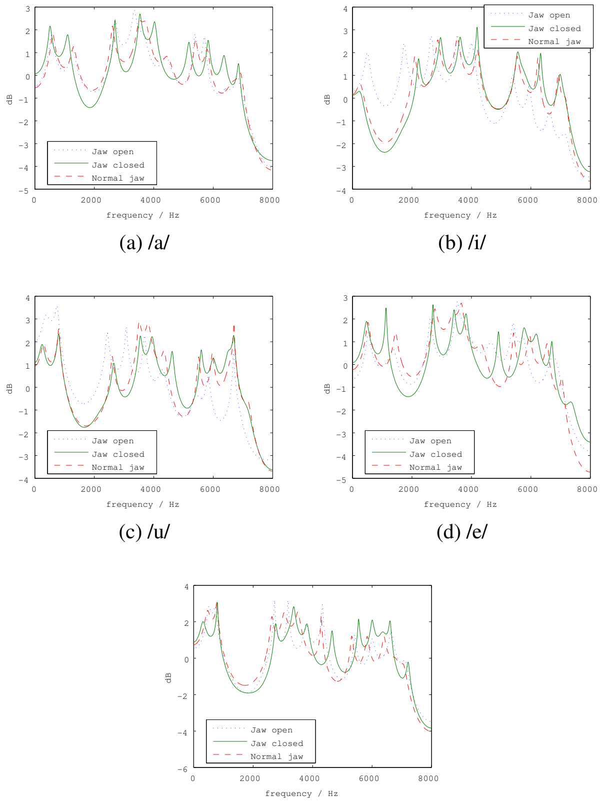


(a) wide open

(b) nearly closed

Figure 3: Running spectrum for /aga/ with jaw fixed to two extreme positions

experiment, since the sound source is maintained the same for all synthetic sounds, the different is resulted from the different positions of the jaw. The spectrum envelops for the synthesized sounds with different jaw positions are shown in figure 4. The gap between the upper and lower jaw is 2.3 cm in the open-jaw cases, and 0.3 cm in closed-jaw cases. From this, we can see that compared with open-jaw case, the closed-jaw cases have a frequency emphasize around the 4 KHz, and it is clearly observed for the vowels /a/ and /e/, because they got a new formant at 4 kHz. Also, for /i/ the power for the existing formant increased. As the three of them takes most of the vowel appearances in the speech corpus, the significant change in these vowels would affect the result a lot. The affection of jaw position is different for different vowels, depending on the normal articulator configuration of the vowel. This result is consistent with the observation from the emotional speech corpus, that when people speaking with their jaw not opening large enough, like in the cold anger emotional state, an emphasis would be generated in certain frequency area of the spectrum.



(a) /a/

(b) /i/

(c) /u/

(d) /e/

(e) /o/

Figure 4: The spectrum envelops for the vowels with the jaw in different positions

For hot anger, the different speaking effort is simulated by controlling the subglottal pressure, the fundamental frequency and the open quotient of the glottal air flow. Figure 5 shows spectrum envelop for the vowel /a/ with different speaking efforts, after they are normalized to have similar integrity. It is obvious that when people speak with more effort, the higher frequency part gets more energy. The analysis results for other vowels are the same as /a/. This is also consistent with the result from the analysis result from the emotional speech database.

As the configuration of the vocal tract is kept the same for the same vowel in this experiment, the transfer function of the vocal tract would be the same. The difference of the spectrum envelop comes from the sound source. Figure 6 shows the glottal air flow with different speaking effort used in the experiment. The spectrum is calculated after the air flows are

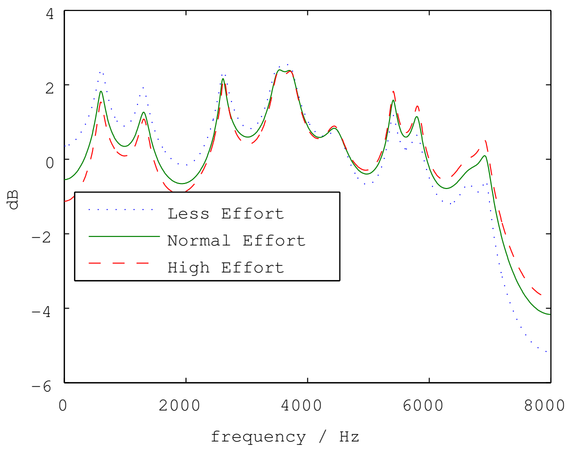
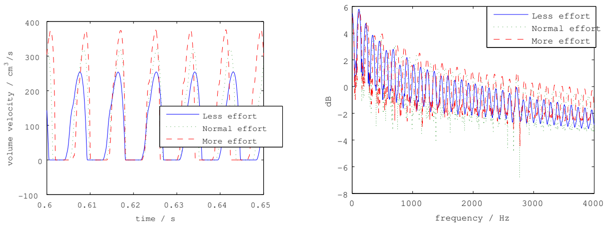


Figure 5: Normalized spectrum envelop with different speaking effort



(a) subglottal air flow

(b) spectrum after normalization

Figure 6: Glottal air flow and its spectrum with different speaking effort

normalized to have similar integrity. With more effort when speaking, the waveform of the glottal wave flow would get sharper, carrying more energy in the high frequency part of the spectrum.

4. Conclusion

With the help a physiological articulatory model, we investigated the relation between certain articulatory configurations and the acoustic features. Fixing the jaw by a gnashing position would generate an spectrum emphasis around 4 kHz, while the extra energy used when speaking would change the spectrum slope, resulting in more energy in the high frequency part of the spectrum. In cold anger and hot anger emotional states, these two kinds of control strategy are used respectively. They both resulted in more energy in the high frequency part of the spectrum.

From the above research, it is confirmed that human beings use a special control strategy, in either phonation or articulation, to generate emotional speech. The control strategy is reflected by the acoustic features of the speech, with which the listeners can understand the emotional state of the speaker.

The physiological articulatory model makes it possible to investigate the relationship between articulatory and acoustic features, which is a great help in the research of human speech generation mechanism and in spotting the spectrum features in emotional speech.

5. Acknowledgement

This study is supported by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC) of Japan, National Natural Science Foundation of China (60805008, 90820304), the National Basic Research Program of China (973 Program) (No. 2006CB303101) and the National High Technology Research and Development Program(“863”Program) of China (No. 2007AA01Z198) .

References

[1] Jianwu Dang and Kiyoshi Honda. A physiological articulatory model for simulating speech production process. *Acoustical Science and Technology*, 22(6):415–425, 2001.

[2] Jianwu Dang and Kiyoshi Honda. Construction and control of a physiological articulatory model. *The Journal of the Acoustical Society of America*, 115(2):853–870, 2 2004.

[3] Shin’ichi Ito, Jianwu Dang, and Masato Akagi. Investigation of the acoustic features of emotional speech using physiological articulatory model. In *International Congress on Acoustics*, volume III, pages 2225–2226, 2004.

[4] Shinji Maeda. A digital simulation method of the vocal-tract system. *Speech Communication*, 1(3–4):199–229, 1982.

[5] Kikuo Maekawa and Takayuki Kagomiya. Influence of paralinguistic information on segmental articulation. In *The Proceedings of the 6th International Conference on Spoken Language Processing (ICASLP2000)*, volume II, pages 349–352, Beijing, China, 2000.

[6] Xiyu Wu, Yongxin Wang, and Jianwu Dang. Investigation of speech production using a 3d physiological articulatory model. In *The Acoustic Society of Japan 2009 Autumn Meeting*, Koriyama, Japan, 2009.