Investigation of the acoustic features of emotional speech using a physiological articulatory model

Shin'ichi Ito, Jianwu Dang, and Masato Akagi.

School of Information Science, Japan Advanced Institute of Science and Technology
{s-itou, jdang, Akagi}@jaist.ac.jp

Abstract

Processing emotional speech is an important issue for speech information science and there are many studies working on this issue. However, we still have no clear knowledge to answer what are the crucial acoustic features for emotional speech, except the fundamental frequency, and how human manipulate their speech organs to generate emotional speech. In this study, we investigate the acoustic features concerned with emotional speech using a physiological articulatory model. The articulatory model is used to simulate human behaviors on production of emotional speech. The speech sound synthesized by the model is used to detect changes in acoustic features when articulatory factors concerned with emotion generation are introduced into the model. Then, the crucial acoustic parameters are identified using a listening test.

1. Introduction

Information involved in spoken language can be subcategorized into three fields: linguistic information concerning with syntaxes and phonetic information; paralinguistic information expressing speaker's intention and attitude; and non-linguistic information that reflects personalities, genders, and emotions. The emotion is one of the important information in human speech communication. Note that the "emotional speech" dealt with in this paper means that the speech can be perceived as a sound with a certain emotion by listeners.

For years, a number of studies have worked on emotional speech. In processing emotional speech, the majority of the studies focused on the acoustic characteristics. So far, the fundamental frequency (F0) is considered as the most important parameter for emotional speech. For some emotion expressions of speech, however, we have difficulties to subcategorize them using the currently known acoustic features. For example, for the normal speech and "cold anger", we can recognize them by our hearing system clearly but we could not distinguish them by the known acoustic features.

There may be some important acoustic features, except the F0, used in human perception processing. However, we do not have a clear knowledge on this issue. The questions here are: Do we neglect some important features for emotion speech? Is there any important parameter for emotional speech we have not discovered yet? Our study begins with these questions.

2. Articulatory Movements *vs* Emotional Speech

When finding a solution for the above questions, the idea first coming to our mind is how human produce emotion speech. In other word, when we express our emotion in speech, how do we manipulate our speech organs? Everyone has experiences in this processing. If we can teach an articulatory model to do as human being does, we may find some solutions from such model simulations.

2.1. Articulatory analysis of emotional speech

In the earlier part, we mentioned the difficulty to distinguish a normal speech and a speech with an emotion of "cold anger". However, the experiences tell us that the articulations for those speech styles have significant differences, in which the jaw is almost fix by gritting the teeth together during the production of "cold anger" speech while the jaw moves freely during a normal speech. Therefore, such articulatory information will be useful for processing emotion speech.

Some studies have endeavored in this direction. For example, Maekawa *et al.* recorded acoustical signal and articulatory movements simultaneously and analyzed the displacement of the speech organs when the speaker intend to express their emotion with speech [2]. They found that for the different emotional speech, the displacement of the speech organs showed different tendencies in the articulatory space. Focusing on the tongue dorsum point, for instance, a number of distinguishing clusters distributed in the anterior-posterior direction. Roughly speaking, the distribution of the articulatory point from the anterior to posterior direction corresponds to the speech with suspicion, disappointment, neutral, and admiration emotions.

2.2. An articulatory model for this study

To lean emotion speech processing from human as mentioned above, it requires a physiological model to faithfully realize the human mechanism. We have developed a physiological articulatory model that can meet this purpose [1].

The physiological model replicates the human speech organs in morphological level and physiological level. The model was constructed based on the volumetric MRI data, which consists of the tongue, jaw, velum, lips, and the vocal tract wall. 11 muscles are involved in the tongue and eight muscles in the jaw complex. A target-based control method is used to generate muscle forces and drives the speech organs towards a given articulatory target. A dynamical vocal tract shape and a series of area functions of the vocal tract are generated based on the muscle-driving model movement. The speech sound is produced using the transmission line model.

2.3. Verification of the articulatory model

Before applying the model in investigation of emotional model, we used some simple tasks to verify the articulatory model. Here, we set the tongue dorsum in different positions and varied the open area of the lips. Figure 1 shows the spectra of /a/ when the open area of the mouth is changed, where *a* indicates the reference, *b* and *c* correspond to the closing and opening actions. Figure 2 shows the spectra of /a/ as the tongue moves in anterior-posterior direction, where *a* indicates the reference, *b* and *c* correspond to forward and backward movements. The results showed that the first formant (F1) becomes higher as a mouth opens wider, and the second formant (F2) increases with a forward movement of the tongue. The results are consistent with the common sense. Since the model can give a good performance for both acoustics and articulation, it is quantified for our research purpose.

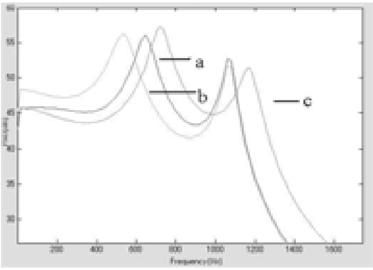

Figure 1. The F1 of /a/ increases as the open area of the mouth increase

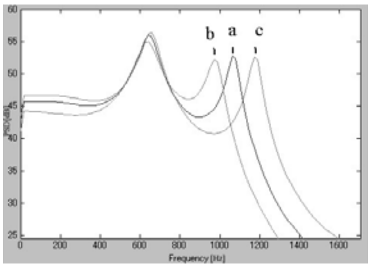

Figure 2. The F2 of /a/ gets higher as the tongue dorsum moves forward.

3. Emotional Speech Synthesis Based on Articulatory Model

The purpose of our synthesis attempts to discover some acoustic features, besides F0, which are important for emotional speech. For this purpose, we focus on the function of the articulators such as movements of the tongue, lips and the jaw, but not change the common acoustical parameters such as F0 and amplitude as possible. Our target is to produce a speech sound with "sad" emotion because there is no significant change in F0. The syllable /ra/ severed in the experiment. First, we generated a sound of /ra/ with a normal position based on the X-ray microbeam data [3]. The sound is defined as the reference sound. To produce a sad sound, the target for the tongue dorsum is moved forward about 1.0cm from the normal position for the vowel /a/. Figure 3 shows the spectrum obtained under the two conditions for /a/. One can see that for the sad speech the F2 increases while the F3 decreases. The acoustic characteristics and articulatory movements show the same tendency as the observation [2]. An informal listening test showed that the listeners can perceive the intended "emotion".

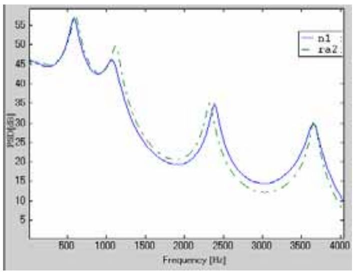

Figure 3. The spectrum of /a/ in /ra/ in normal speech (sold line) and with "sad" emotion (dashed line).

4. Summary

In this study we attempted to discover some new features that deeply concerned with emotional speech. We have not obtained definite results yet. However, the primary results showed that it is a feasible way to use an articulatory model in the study on emotional speech processing.

5. References

[1] Dang, J. and Honda, K. "Construction and control of a physiological articulatory model," J. Acoust. Soc. Am. 115 (3), (2004), in press

[2] Kikuo Maekawa and Takayuki Kagomiya, "Influence of Paralinguistic Information on Segmental Articulation," CREST/JST, 2002

[3] Hashi, M., Westbury, J. & Honda, K. (1998) "Vowel posture normalization," J. Acous. Soc. Am. 104, 2426-2437.