The Eighteenth International Symposium on Artificial Life and Robotics 2013 (AROB 18th ’13),
Daejeon Convention Center, Daejeon, Korea, January 30-February 1, 2013

# Speech synthesis of emotions using vowel features of a speaker

K. Boku, T. Asada, Y. Yoshitomi, and M. Tabuse

Graduate School of Life and Environmental Sciences, Kyoto Prefectural University,
1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan
E-mail: {boku, t_asada}@mei.kpu.ac.jp, {yoshitomi, tabuse}@kpu.ac.jp

**Abstract:** Recently, methods for adding emotion to synthetic speech have received considerable attention in the field of speech synthesis research. We previously proposed a case-based method for generating emotional synthetic speech by exploiting the characteristics of the maximum amplitude and the utterance time of vowels, and the fundamental frequency of emotional speech. In the present study, we propose a method in which our reported method is further improved by controlling the fundamental frequency of emotional synthetic speech. As an initial investigation, we adopted the utterance of a Japanese name that is semantically neutral. By using the proposed method, emotional synthetic speech made from the emotional speech of one male subject was discriminable with a mean accuracy of 90.0% when 18 subjects listened to the emotional synthetic utterances of “angry,” “happy,” “neutral,” “sad,” or “surprised” when the utterance was the Japanese name “Taro.”

**Keywords:** Emotional speech, Feature parameter, Synthetic speech, Emotional synthetic speech, Vowel

## 1 INTRODUCTION

Recently, methods for adding emotions to synthetic speech have received considerable attention in the field of speech synthesis research [1]-[8]. To generate emotional synthetic speech, it is necessary to control the prosodic features of the utterances. Natural language is mainly composed of vowels and consonants. The Japanese language has five vowels. A vowel has a more dominant impact on the listener's impression than does a consonant, primarily because a vowel has a longer utterance time and a larger amplitude in comparison with a consonant. We previously proposed a case-based method for generating emotional synthetic speech by exploiting the characteristics of the maximum amplitude and the utterance time of the vowels, as obtained by using a speech recognition system, and the fundamental frequency of emotional speech [9].

In the present study, we propose a method in which the method reported in [9] is further improved by controlling the fundamental frequency of emotional synthetic speech. The advantage of our study over reported research [1]-[8] is usage of the vowel feature in emotional speech to generate emotional synthetic speech.

## 2 PROPOSED METHOD

In the first stage, we obtain audio data of emotional speech as a WAV file when a subject speaks with each of the intentional emotions of “angry,” “happy,” “neutral,” “sad,” and “surprised.” Then, for each kind of emotional speech, we measure the time of each vowel utterance and the value of the maximum amplitude of the waveform while speaking the vowel, and the fundamental frequency of emotional speech.

In the second stage, we synthesize the phoneme sequence uttered by the subject. This stage consists of the following five steps.

***Step 1***: For a vowel with a consonant appearing just before it in synthetic speech with neutral emotion, the total phonation duration time of the vowel and the consonant is transformed into the time for speech with neutral emotion by the human subject. The synthetic speech obtained by this processing is hereinafter called “neutral synthetic speech.”
***Step 2***: For a vowel with a consonant appearing just before it in synthetic speech with one of the intentional emotions of “angry,” “happy,” “sad,” and “surprised,” the total phonation duration time of the vowel and consonant is set as the value whose ratio to that in neutral synthetic speech is equal to the ratio of the phonation duration time of the vowel in emotional speech to the phonation duration time of the vowel in neutral speech.
***Step 3***: The fundamental frequency of synthetic speech, obtained by the processing up to Step 2, is initially adjusted based on the fundamental frequency of the emotional speech.
***Step 4***: For a vowel with a consonant appearing just before it in synthetic speech obtained by the processing up to Step 3, the amplitudes are transformed into final values by twice multiplying the ratio $\left(Max_{em}/Max_{ne}\right)$, where $Max_{em}$ and $Max_{ne}$ denote the maximum amplitude of the vowel in emotional speech and that in neutral speech, respectively. The synthetic speech obtained by the processing of Steps 1 to 4 is hereinafter called “emotional synthetic speech.”

***Step 5***: The fundamental frequency of emotional synthetic speech, obtained by the processing up to Step 4, is further adjusted based on the fundamental frequency of the emotional speech.

If no consonant appears, the process described in Steps 1 to 5 applies to just the vowel. In the present study, the processing described in Step 5 is added to the method reported in [9].

## 3 EXPERIMENTS

### 3.1 Condition

We used a speech recognition system named Julius [10] to save the timing positions of the start of speech, and the first and last vowels. A male subject (Subject A) in his 50s spoke the semantically neutral utterance of the Japanese first name "Taro" with each of the intentional emotions of "angry," "happy," "neutral," "sad," and "surprised." His audio data were recorded as WAV files. We measured the utterance time of the first vowel and the maximum absolute value of the amplitude of the waveform while speaking the first vowel. For the last vowel, we also measured the utterance time and the maximum absolute value of the amplitude of the waveform. Tables 1 and 2 show the phonation time and the maximum amplitude, respectively, of each vowel in each emotion category as spoken by the subject.

**Table 1.** Phonation time of vowels as spoken by the subject [9]

| Emotion category | Phonation time (s) | | Normalized phonation time (Ratio to the value of "Neutral") | |
|---|---|---|---|---|
| | /a/ | /o/ | /a/ | /o/ |
| Angry | 0.030 | 0.090 | 1.000 | 0.310 |
| Happy | 0.100 | 0.240 | 3.333 | 0.828 |
| Neutral | 0.030 | 0.290 | 1.000 | 1.000 |
| Sad | 0.100 | 0.140 | 3.333 | 0.483 |
| Surprised | 0.050 | 0.200 | 1.667 | 0.690 |

**Table 2.** Maximum amplitude of vowels as spoken by the subject [9]

| Emotion category | Maximum amplitude | | Normalized maximum amplitude (Ratio to the value of "Neutral") | |
|---|---|---|---|---|
| | /a/ | /o/ | /a/ | /o/ |
| Angry | 1216 | 459 | 1.332 | 0.630 |
| Happy | 1904 | 1055 | 2.085 | 1.447 |
| Neutral | 913 | 729 | 1.000 | 1.000 |
| Sad | 587 | 295 | 0.643 | 0.405 |
| Surprised | 1408 | 1256 | 1.542 | 1.723 |

We performed a principal component analysis (PCA) to reveal the prosodic characteristics of "angry," "happy," "neutral," "sad," and "surprised" in emotional speech by using the normalized utterance time and the normalized maximum amplitude of the first and last vowels as the feature parameters. Here, normalization of the utterance time and the maximum amplitude was performed by setting the mean value for the five emotions to zero and setting the standard deviation for each of the five emotions to one. Fig. 1 shows the feature vector space expressed by the first and second components obtained by PCA for the five kinds of emotional speech for /taro/.

Then, the fundamental frequency of emotional speech with each of the intentional emotions of "angry," "happy," "neutral," "sad," and "surprised" was measured for the Japanese first name "Taro." The fundamental frequency was selected because it is one of the best-known feature parameters for speech. As shown in Fig. 2, each emotion had the characteristic time-dependence of the fundamental frequency.

Voice Sommelier Neo (premium version; Hitachi Business Solution Co., Ltd., Yokohama, Japan) [11] was

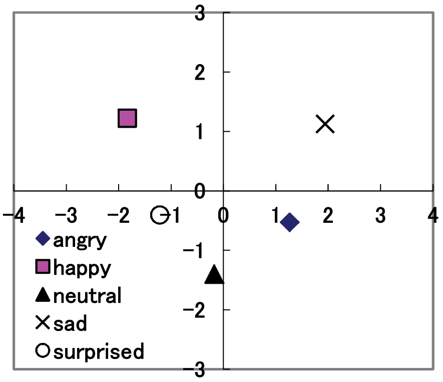

**Fig. 1.** Feature space of the first (horizontal) and second (vertical) components for /taro/ [9]

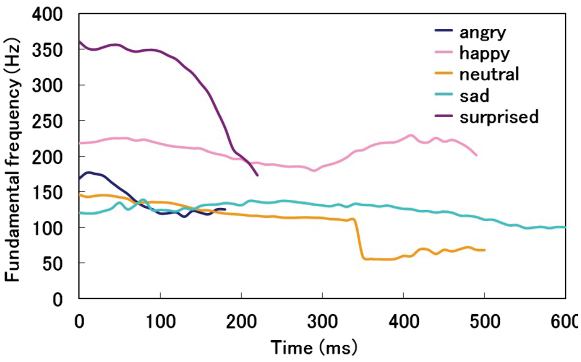

**Fig. 2.** Fundamental frequencies of waveforms of Subject A while speaking "Taro"

used as the speech synthesizer for Steps 1 to 3 in Section 2. For conversion of the amplitude of each vowel and consonant described in Step 4 in Section 2, a digital audio editor was used. Then, the method [12] using resampling and a correlation function were used for Step 5 in Section 2.

When we applied the method described in Section 2 to the above case, the mode of Male 1 (bright voice) in Voice Sommelier Neo was used. In this case, each vowel was /a/ and /o/, and then the vowel and the consonant just before the vowel was /ta/ and /ro/.

The emotional synthetic speech without adding the processing described in Step 5 of Section 2 is hereinafter called "emotional synthetic speech No. 1," whereas the emotional synthetic speech with adding the processing described in Step 5 of Section 2 is hereinafter called "emotional synthetic speech No. 2." The emotional synthetic speech No. 1 was made only by the method reported in [9] and used for the evaluation of Step 5 of Section 2, which is the new processing in the proposed method.

The 18 subjects participating in the experiments consisted of the following: Subjects A and B, two males in their 50s; Subject C, one male in his 30s; Subjects D, E, F, G, H, I, J, K, L, M and N, 11 males in their 20s; Subjects O, P, Q and R, four females in their 20s. The subjects made a judgment of the emotional category after listening to five utterances one by one in the following order: emotional speech by Subject A, emotional synthetic speech No. 1, emotional speech by Subject A, and emotional synthetic speech No. 2.

### 3.2 Results and discussion

Figs. 3 and 4 show the fundamental frequencies of emotional synthetic speech Nos. 1 and 2, respectively. The characteristics of each type of emotional speech shown in Fig. 2 are more precisely reflected in Fig. 4, which was obtained by the proposed method, than they were in Fig. 3, which was obtained by the method reported in [9]. As illustrated in Fig. 5, differences among the emotional speech waveforms were observed. To some extent, the differences of the waveforms were also reflected in each emotional synthetic speech. Voice Sommelier Neo used in Step 3 in Section 2 has some restrictions in controlling the frequency of synthetic speech, so it was difficult to adjust the fundamental frequency adequately. The differences between the waveforms of the emotional speech waveform of "neutral" and the synthetic speech waveform of "neutral" are shown in Fig. 5. As expected, the emotional synthetic speech inevitably had some waveform differences in comparison with the corresponding emotional speech.

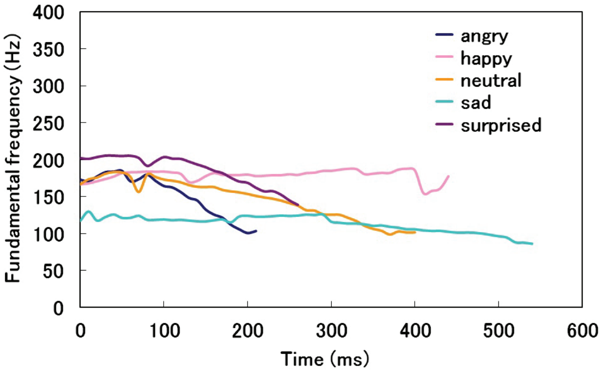

**Fig. 3.** Fundamental frequencies of waveforms of emotional synthetic speech No. 1 of "Taro"

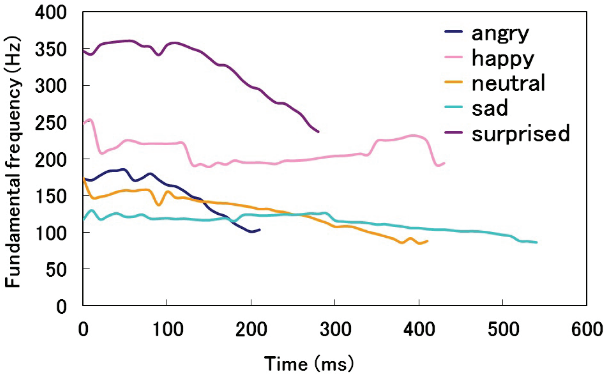

**Fig. 4.** Fundamental frequencies of waveforms of emotional synthetic speech No. 2 of "Taro"

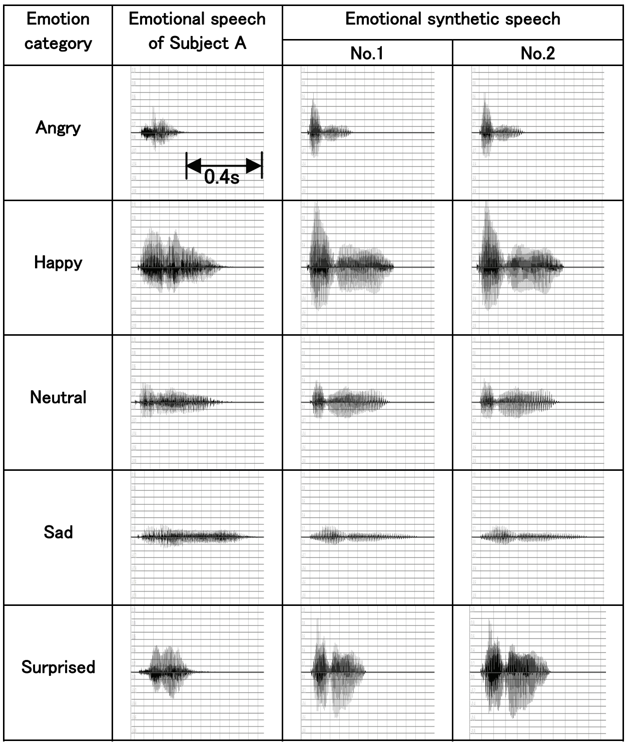

**Fig. 5.** Waveforms of emotional speech of Subject A and emotional synthetic speech Nos. 1 and 2

Table 3 shows the results of the subjective evaluations. In Table 3, the results for emotional speech were calculated as an average of the values obtained in two sets of listening by all 18 subjects. As illustrated in Table 3, the mean accuracy of emotional speech, emotional synthetic speech No. 1, and emotional synthetic speech No. 2 was 95.0%, 70.0%, and 90.0%, respectively. The advantage of emotional synthetic speech No. 2 over No. 1 suggests that further adjustment of fundamental frequency in the proposed method made a much clearer impression on the subjects for emotional synthetic speech.

**Table 3.** Results of the subjective evaluation

(1) Emotional speech

| | | Input | | | | |
|---|---|---|---|---|---|---|
| | | Angry | Happy | Neutral | Sad | Surprised |
| Recognition | Angry | 97.2 | 0 | 0 | 2.8 | 0 |
| | Happy | 0 | 94.4 | 0 | 0 | 5.6 |
| | Neutral | 0 | 2.8 | 94.4 | 2.8 | 0 |
| | Sad | 0 | 0 | 5.6 | 94.4 | 0 |
| | Surprised | 2.8 | 2.8 | 0 | 0 | 94.4 |

(%)

(2) Emotional synthetic speech No.1

| | | Input | | | | |
|---|---|---|---|---|---|---|
| | | Angry | Happy | Neutral | Sad | Surprised |
| Recognition | Angry | 83.3 | 0 | 0 | 0 | 22.2 |
| | Happy | 0 | 50.0 | 38.8 | 0 | 5.6 |
| | Neutral | 0 | 44.4 | 55.6 | 11.1 | 0 |
| | Sad | 0 | 0 | 5.6 | 88.9 | 0 |
| | Surprised | 16.7 | 5.6 | 0 | 0 | 72.2 |

(%)

(3) Emotional synthetic speech No.2

| | | Input | | | | |
|---|---|---|---|---|---|---|
| | | Angry | Happy | Neutral | Sad | Surprised |
| Recognition | Angry | 88.8 | 5.6 | 0 | 0 | 11.1 |
| | Happy | 5.6 | 88.8 | 0 | 0 | 0 |
| | Neutral | 0 | 5.6 | 94.4 | 11.1 | 0 |
| | Sad | 0 | 0 | 5.6 | 88.9 | 0 |
| | Surprised | 5.6 | 0 | 0 | 0 | 88.9 |

(%)

## 4 CONCLUSION

We previously proposed a case-based method for generating emotional synthetic speech by exploiting the characteristics of the maximum amplitude and the utterance time of the vowels, as obtained by using a speech recognition system, and the fundamental frequency of emotional speech. In the present study, we propose a method in which our reported method is further improved by controlling the fundamental frequency of emotional synthetic speech. By using the proposed method, emotional synthetic speech made from the emotional utterances of one male subject was discriminable with a mean accuracy of 90.0% when 18 subjects listened to one of the emotional synthetic speech utterances of "angry," "happy," "neutral," "sad," or "surprised" for the Japanese name "Taro."

### Acknowledgments

We would like to thank all the participants who cooperated with us in the experiments.

## REFERENCES

[1] Donna E (2005), Expressive speech: Production, perception and application to speech synthesis. Acoustical Science and Technology, 26(4):317–325

[2] Iida A, Iga S, Higuchi F, et al (2000), A prototype of a speech synthesis system with emotion for assisting communication (in Japanese). Transaction of Human Interface Society, 2(2): 63-70

[3] Katae N, Kimura S (2000), An effect of voice quality and control in emotional speech synthesis (in Japanese). Proceedings of the autumn meeting the Acoustical Society of Japan: Vol. 2. pp.187-188

[4] Moriyama T, Mori S, Ozawa S (2009), A synthesis method of emotional speech using subspace constraints in prosody (in Japanese). Transaction of Information Processing Society of Japan, 50(3): 1181-1191

[5] Murray IR, Arnott JL (1993), Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Journal of the Acoustical Society of America, 93(2):1097–1108

[6] Murray IR, Edgington MD, Campion D, et al (2000), Rule-based emotion synthesis using concatenated speech. Proceedings of Speech and Emotion, ISCA Tutorial and Research Workshop, Newcastle, Northern Ireland, UK pp.173–177

[7] Ogata S, Yotsukura T, Morishima S (2000), Voice conversion to append emotional impression by controlling articulation information (in Japanese). IEICE technical report, human information processing, 99(582):53-58

[8] Schröder M (2001), Emotional speech synthesis – A review. In: Dalsgaard P, Lindberg B, Benner H (eds), Proceedings of 7th European Conference on Speech Communication and Technology, Aalborg, Kommunik Grafiske Losninger A/S, Vol.1, pp.561–564

[9] Boku K, Asada T, Yoshitomi Y et al (2012), In: Roger L (ed), Software engineering, artificial intelligence, networking, and parallel/distributed computing 2012, Springer, Berlin, Germany, pp.129-141

[10]Julius development team (2011), Open-source large vocabulary CSR engine Julius. Retrieved June 23, 2012, from http://julius.sourceforge.jp/en_index.php?q=index-en.html

[11]Hitachi Business Solution Co. Ltd. (2012), Voice Sommelier Neo. Retrieved June 23, 2012, from http://hitachi-business.com/products/package/sound /voice/index.html

[12]Aoki N (2008), Sound programming in C (in Japanese). Ohmsha, Tokyo, Japan, pp.141-160