Journal of Robotics, Networking and Artificial Life
Vol. 7(2); September (2020), pp. 107–110
DOI: https://doi.org/10.2991/jrnal.k.200528.007; ISSN 2405-9021; eISSN 2352-6386
https://www.atlantis-press.com/journals/jrnal

Research Article

# Speech Synthesis of Emotions in a Sentence using Vowel Features

Rintaro Makino[1], Yasunari Yoshitomi[2,*], Taro Asada[2], Masayoshi Tabuse[2]

[1]*UX Solution Department UX Division, Customer Platform Promotion Division, SoftBank Corp., 1-9-1 Higashi-shimbashi, Minato-ku, Tokyo 105-7317, Japan*
[2]*Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, 1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan*

**ARTICLE INFO**

*Article History*
Received 14 November 2019
Accepted 28 April 2020

*Keywords*
Emotional speech
feature parameter
emotional synthetic speech
vowel
sentence

**ABSTRACT**

We previously proposed a method for adding emotions to synthetic speech using the vowel features of a speaker. For the initial investigation in this earlier study, we used utterances of Japanese names to demonstrate the method. In the present study, we use the proposed method to construct emotional synthetic speech for a sentence formed from the emotional speech of a single male subject and produce results that are discriminable with good accuracy.



## 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 larger amplitude in comparison to a consonant. We previously proposed a case-based method for generating emotional synthetic speech by exploiting the characteristics of maximum amplitude and utterance time for vowels, as obtained by using a speech recognition system, together with the fundamental frequency of emotional speech [9].

In an earlier study [10], we proposed an approach that further improved the method described in Boku et al. [9] by controlling the fundamental frequency of the emotional synthetic speech. The advantage of this earlier study [10] over prior research [1–8] was the usage of the vowel feature in emotional speech to generate synthetic emotional speech. In the previous study [10], the speech included only Japanese names.

In the present study, we apply our previously proposed method [10] to creating emotional synthetic speech in sentence form, with new treatments for (1) smoothing the fundamental frequencies of sequential vowels and (2) suitably jointing the synthetic phonemes.

*Corresponding author. Email: yoshitomi@kpu.ac.jp

## 2. PROPOSED METHOD

In the first stage, we obtain audio data for emotional speech recorded as a WAV file. The subject is asked to speak in a way that reflects various emotional states: "angry," "happy," "neutral," "sad," and "surprised." Then, for each type of emotional speech, we measure the time of each vowel utterance and the value of the maximum amplitude of the waveform while the vowel is being spoken, as well as the fundamental frequency of the emotional speech [10].

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

**Step 1:** For a vowel with a consonant appearing just before it in the 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" [10].

**Step 2:** For a vowel with a consonant appearing just before it in the synthetic speech reflecting one of the intentional emotions ("angry," "happy," "sad," or "surprised"), the total phonation duration time of the vowel and consonant is set so that the ratio of this total to the corresponding total in the neutral synthetic speech is equal to the ratio of the phonation duration time of the vowel in the emotional speech to the phonation duration time of the vowel in the neutral speech [10].

**Step 3:** The fundamental frequency of the synthetic speech obtained by the processing conducted up through Step 2 is initially adjusted based on the fundamental frequency of the emotional speech [10].

**Step 4:** For a vowel with a consonant appearing just before it in the synthetic speech obtained by the processing conducted up through

Step 3, the fundamental frequency is transformed into the average of three values of the sequential vowels that include it. For the first or last vowel in the sentence, two sets of fundamental frequencies are used for averaging.

**Step 5:** Amplitudes are transformed into final values by twice multiplying the ratio ($Max_{em}/Max_{ne}$), where $Max_{em}$ and $Max_{ne}$ denote the maximum amplitude of the vowel in emotional speech and the amplitude of the vowel in neutral speech, respectively. The synthetic speech obtained by the processing up through Step 5 is hereinafter called "emotional synthetic speech."

**Step 6:** The fundamental frequency of the emotional synthetic speech obtained by the processing conducted up through Step 5 is further adjusted based on the fundamental frequency of the emotional speech.

**Step 7:** The synthetic speech for a sentence is generated by jointing all the synthetic speech phonemes constructing the sentence and adjusting the speech length to the phonation duration time of the emotional speech obtained in advance.

If no consonant appears before a vowel, the process described in Steps 1 through 6 applies only to the vowel.

# 3. EXPERIMENT

## 3.1. Condition

We used the speech recognition system known as Julius [11] to recognize the vowels and save the time position of the start and end of the vowels in each utterance. One male subject (Subject A) in his 50s was asked to speak the Japanese first names listed in Table 1 in a way that reflected various emotional states: "angry," "happy," "neutral," "sad," and "surprised." The audio data were recorded as WAV files. In preparation for generating the emotional synthetic speech, we measured the utterance time of the vowel, the maximum amplitude of the waveform, and the fundamental frequency while the vowel was being spoken. Tables 2–4 show the phonation time average, the maximum amplitude average, and the fundamental frequency average, respectively, for each first vowel in each Japanese name in each emotion category as spoken by Subject A.

Voice Sommelier Neo (premium version; Hitachi Business Solution Co., Ltd., Yokohama, Japan) [12] was used as the speech synthesizer for each of the steps described in Section 2. In applying the method, the Male 1 (bright voice) mode in Voice Sommelier Neo was used. To convert the amplitude of each vowel and consonant described in Step 5 in Section 2, a digital audio editor was used. The method [13] using resampling was then used in Step 6 of Section 2.

To enable comparisons, subject A was asked to speak the sentence, 'このぬいぐるみかわいくない' (in Japanese), which means, 'This stuffed toy is pretty, isn't it?" in a way that reflected each of the intentional emotions. ("Surprised" was not included here, as it was difficult to express in the sentence.) We used the recorded utterances as "emotional speech." We then generated synthetic speech for the same sentence ('このぬいぐるみかわいくない') using the method described in Section 2.

In all, 13 subjects participated in the experiment. These included one male in his 60s, one male in his 50s, one male in his 30s, five males in their 20s, and five females in their 20s. Each of the 13 subjects was asked to judge the emotional state of the speaker (angry, happy, neutral, sad) after listening to four utterances, one at a time, in the following order: the emotional speech by Subject A, the emotional synthetic speech. The 13 subjects were also asked which features were very important in judging the emotion of the sentence. There were seven choices available: three prosodic features, (1) utterance length, (2) height, and (3) volume; and four in-sentence positional features, (4) top, (5) middle, (6) last, and (7) total. For the three prosodic features, multiple answers were allowed, while only one answer was allowed for in-sentence positional features.

## 3.2. Results and Discussion

Table 5 shows the results of the subjective evaluations of the 13 subjects. As indicated, the mean accuracy of the categorizations for the

**Table 1** | Japanese first name used in experiments

| | | First vowel | | | | |
|---|---|---|---|---|---|---|
| | | /a/ | /i/ | /u/ | /e/ | /o/ |
| Last vowel | /a/ | ayaka | shinnya | tsubasa | keita | tomoya |
| | /i/ | kazuki | hikari | yuki | megumi | koji |
| | /u/ | takeru | shigeru | fuyu | megu | noboru |
| | /e/ | kaede | misae | yusuke | keisuke | kozue |
| | /o/ | taro | hiroko | yuto | keiko | tomoko |

**Table 2** | Phonation time average of each first vowel

| | | Emotion | | | | |
|---|---|---|---|---|---|---|
| | | Angry | Happy | Neutral | Sad | Surprised |
| First vowel | /a/ | 62 | 110 | 72 | 112 | 58 |
| | /i/ | 44 | 74 | 68 | 86 | 58 |
| | /u/ | 48 | 120 | 130 | 86 | 44 |
| | /e/ | 70 | 126 | 136 | 182 | 84 |
| | /o/ | 70 | 104 | 118 | 176 | 84 |

(ms)

**Table 3** | Maximum amplitude average of each first vowel

| | | Emotion | | | | |
|---|---|---|---|---|---|---|
| | | Angry | Happy | Neutral | Sad | Surprised |
| First vowel | /a/ | 1340 | 1714 | 714 | 573 | 1346 |
| | /i/ | 362 | 385 | 287 | 199 | 400 |
| | /u/ | 658 | 509 | 438 | 298 | 1017 |
| | /e/ | 816 | 1079 | 794 | 575 | 1165 |
| | /o/ | 748 | 1262 | 838 | 646 | 1479 |

**Table 4** | Fundamental frequency average of each first vowel

| | | Emotion | | | | |
|---|---|---|---|---|---|---|
| | | Angry | Happy | Neutral | Sad | Surprised |
| First vowel | /a/ | 141 | 183 | 131 | 165 | 284 |
| | /i/ | 114 | 192 | 134 | 191 | 286 |
| | /u/ | 159 | 217 | 131 | 165 | 284 |
| | /e/ | 122 | 186 | 178 | 229 | 276 |
| | /o/ | 139 | 186 | 170 | 210 | 228 |

(Hz)

emotional speech was 100%; for the emotional synthetic speech, the mean accuracy was 78.9%. The accuracy for the "happy" synthetic speech was highest, while that of "neutral" was lowest. Table 6 shows the results of the questions regarding the importance of various features in judging the emotional category. As shown, among the three prosodic characteristics, the "height of the voice" had the most influence, while among the four sentence positions, the "last position" appears to have influenced the subjects the most. Figure 1 shows the waveforms of the emotional speech and the emotional synthetic speech. As illustrated in the figure, the waveform associated with the "happy" synthetic speech appears to be the most similar to its emotional speech counterpart among the waveforms of the four emotional synthetic speech types.

Tables 7 and 8 show the results of the subjective evaluations for the names "Taro" and "Hiroko," respectively, as reported in our previous study [10]. In these two tables, the results for the emotional speech were calculated as the average of the values obtained in two sets of listening by all 18 subjects. Based on the values in Table 7, for "Taro," the mean accuracy for the emotional speech categorizations using the previously proposed method was 95.0%, while for the emotional synthetic speech, it was 90.0% [10]. For "Hiroko," the mean accuracy for the emotional speech categorizations was 96.7%, while for the emotional synthetic speech, it was 77.8% (Table 8) [10]. Combining the results for both "Taro" and "Hiroko," the average accuracy for the emotional speech categorizations was 95.9%, while the average accuracy for the emotional synthetic speech was 83.9%.

A comparison of the combined 83.9% average accuracy reported for the emotional synthetic speech categorizations in the previous study (Tables 7 and 8) [10] and the lower 78.9% average accuracy in the present study [Table 5 (2)] suggests that it may be more difficult to effectively generate emotional synthetic speech in a sentence rather than for a first name (It should be noted that there were four emotion categories in the more recent experiment versus five in the previous study [10]. As mentioned earlier, "surprised" was omitted in the current study.).

**Table 5** | Results of subjective evaluations

| | | Input | | | |
|---|---|---|---|---|---|
| | | **Angry** | **Happy** | **Neutral** | **Sad** |
| | | **(1) Emotional speech** | | | |
| Recognition | Angry | **100** | 0 | 0 | 0 |
| | Happy | 0 | **100** | 0 | 0 |
| | Neutral | 0 | 0 | **100** | 0 |
| | Sad | 0 | 0 | 0 | **100** |
| | | **(2) Emotional synthetic speech** | | | |
| Recognition | Angry | **76.9** | 7.7 | 15.4 | 0 |
| | Happy | 0 | **92.3** | 0 | 15.4 |
| | Neutral | 23.1 | 0 | **69.2** | 7.7 |
| | Sad | 0 | 0 | 15.4 | **76.9** |

(%)

**Table 6** | Results of questions regarding the importance of various factors in judging the emotion category

| | **Voice** | | | **Sentence position** | | | |
|---|---|---|---|---|---|---|---|
| | **Length** | **Height** | **Volume** | **Top** | **Middle** | **Last** | **Total** |
| Number of vote | 7 | 11 | 9 | 0 | 0 | 8 | 5 |
| Ratio (%) | 54 | 85 | 69 | 0 | 0 | 62 | 38 |

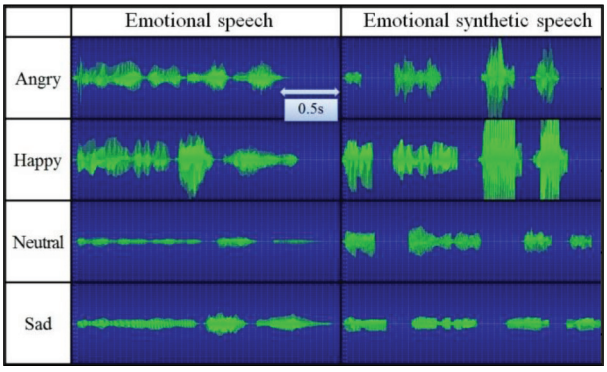


**Figure 1** | Waveform of emotional speech and emotional synthetic speech for the intentional emotions of "angry," "happy," "neutral," and "sad" in the utterance of 'このぬいぐるみかわいくない' (in Japanese).

**Table 7** | Results of the subjective evaluation for "Taro" [10]

| | | Input | | | | |
|---|---|---|---|---|---|---|
| | | **Angry** | **Happy** | **Neutral** | **Sad** | **Surprised** |
| | | **(1) Emotional speech** | | | | |
| 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** | | | | |
| 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** |

(%)

**Table 8** | Results of the subjective evaluation for "Hiroko" [10]

| | | Input | | | | |
|---|---|---|---|---|---|---|
| | | **Angry** | **Happy** | **Neutral** | **Sad** | **Surprised** |
| | | **(1) Emotional speech** | | | | |
| Recognition | Angry | **91.7** | 0 | 0 | 0 | 8.3 |
| | Happy | 0 | **100** | 0 | 0 | 0 |
| | Neutral | 0 | 0 | **100** | 0 | 0 |
| | Sad | 0 | 0 | 0 | **100** | 0 |
| | Surprised | 8.3 | 0 | 0 | 0 | **91.7** |
| | | **(2) Emotional synthetic speech** | | | | |
| Recognition | Angry | **83.3** | 0 | 0 | 0 | 16.6 |
| | Happy | 0 | **66.7** | 0 | 22.2 | 5.6 |
| | Neutral | 0 | 27.8 | **94.4** | 11.1 | 0 |
| | Sad | 0 | 0 | 5.6 | **66.7** | 0 |
| | Surprised | 16.7 | 5.5 | 0 | 0 | **77.8** |

(%)

## 4. CONCLUSION

Recently, methods for adding emotion to synthetic speech have received considerable attention in speech synthesis research. We

had previously proposed a method for speech synthesis with emotions using the vowel features of a speaker [10]. In the present study, we used the proposed method [10] to create emotional synthetic speech in a sentence based on the emotional speech of an individual and found that the results were discriminable with a good accuracy.

## CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

## ACKNOWLEDGMENTS

We would like to thank all those who participated in the experiments described here.

## REFERENCES

[1] D. Erickson, Expressive speech: production, perception and application to speech synthesis, Acoust. Sci. Technol. 26 (2005), 317–325.
[2] A. Iida, S. Iga, F. Higuchi, N. Campbell, M. Yasumura, A speech synthesis system with emotion for assisting communication, Trans. Human Interface Soc. 2 (2000), 63–70 (in Japanese).
[3] N. Katae, S. Kimura, An effect of voice quality and control in emotional speech synthesis, Proceedings of the Autumn Meeting the Acoustical Society of Japan, vol. 2, Iwate, Japan, 2000, pp. 187–188 (in Japanese).
[4] T. Moriyama, S. Mori, S. Ozawa, A synthesis method of emotional speech using subspace constraints in prosody, Trans. Inform. Process. Soc. J. 50 (2009), 1181–1191 (in Japanese).
[5] I.R. Murray, J.L. Arnott, Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J. Acoust. Soc. Am. 93 (1993), 1097–1108.
[6] I.R. Murray, M.D. Edgington, D.M. Campion, J.F. Lynn, Rule-based emotion synthesis using concatenated speech. Proceedings of ISCA Tutorial and Research Workshop on Speech and Emotion, International Speech Communication Association, Newcastle, UK, 2000, pp. 173–177.
[7] S. Ogata, T. Yotsukura, S. Morishima, Voice conversion to append emotional impression by controlling articulation information. IEICE Tech. Rep. Human Inf. Process. 99 (2000), 53–58 (in Japanese).
[8] M. Schröder, Emotional speech synthesis—a review, in: P. Dalsgaard, B. Lindberg, H. Benner, Proceedings of 7th European Conference on Speech Communication and Technology, vol. 1, Kommunik Grafiske Losninger A/S, Aalborg, 2001, pp. 561–564.
[9] K. Boku, T. Asada, Y. Yoshitomi, M. Tabuse, Speech synthesis of emotions using vowel features, in: R. Lee (Eds.), Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing 2012, Springer, Berlin, Heidelberg, 2013, pp. 129–141.
[10] K. Boku, T. Asada, Y. Yoshitomi, M. Tabuse, Speech synthesis of emotions using vowel features of a speaker, J. Artif. Life Robot. 19 (2014), 27–32.
[11] Julius Development Team, Open-source large vocabulary CSR engine Julius, Available from: http://julius.sourceforge.jp/ (accessed December 11, 2019).
[12] Hitachi Business Solution Co. Ltd., Voice Sommelier Neo, Available from: https://www.hitachi-solutions-create.co.jp/solution/voice/ (accessed April 28, 2020).
[13] N. Aoki, Sound programming in C, Ohmsha, Tokyo, 2008, pp. 141–160 (in Japanese).

## AUTHORS INTRODUCTION

**Mr. Rintaro Makino**

He received his B.S. degree from Kyoto Prefectural University in 2018. He works at SoftBank Corp.

**Dr. Taro Asada**

He received his B.S., M.S. and PhD degrees from Kyoto Prefectural University in 2002, 2004 and 2010, respectively. He works as an Associate Professor at the Graduate School of Life and Environmental Sciences of Kyoto Prefectural University. His current research interests are human interface and image processing. He is a member of HIS, IIEEJ.

**Dr. Yasunari Yoshitomi**

He received his B.E, M.E. and PhD degrees from Kyoto University in 1980, 1982 and 1991, respectively. He works as a Professor at the Graduate School of Life and Environmental Sciences of Kyoto Prefectural University. His specialties are applied mathematics and physics, informatics environment, intelligent informatics. He is a member of IEEE, HIS, ORSJ, IPSJ, IEICE, SSJ, JMTA and IIEEJ.

**Dr. Masayoshi Tabuse**

He received his M.S. and PhD degrees from Kobe University in 1985 and 1988 respectively. From June 1992 to March 2003, he had worked in Miyazaki University. Since April 2003, he has been in Kyoto Prefectural University. He works as a Professor at the Graduate School of Life and Environmental Sciences of Kyoto Prefectural University. His current research interests are machine learning, computer vision and natural language processing. He is a member of IPSJ, IEICE and RSJ.