

Research Article

Facial Expression Synthesis System Using Speech Synthesis and Vowel Recognition

Taro Asada¹, Ruka Adachi², Syuhei Takada³, Yasunari Yoshitomi¹, Masayoshi Tabuse¹¹Graduate School of Life and Environmental Sciences, Kyoto Prefectural University, 1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan²Second System Department, Software Service, Inc., 2-6-1 Nishi-Miyahara, Yodogawa-Ku, Osaka, Japan³Resident Department, Seika Town Hall, 70 Kitashiri, Minamiinayazuma, Nishi-Miyahara, Sagara-Gun, Kyoto, Japan

ARTICLE INFO

Article History

Received 14 November 2019

Accepted 20 July 2020

Keywords

MMD Agent

Speech recognition

Vowel recognition

Speech synthesis

ABSTRACT

Herein, we report on the development of a system for agent facial expression generation that uses vowel recognition when generating synthesized speech. The speech is recognized using the Julius high-performance, two-pass large vocabulary continuous speech recognition decoder software system, after which the agent's facial expression is synthesized using preset parameters that depend on each vowel. The agent was created using MikuMikuDanceAgent (MMDAgent), which is a freeware animation program that allows users to create and animate movies with agents.

© 2022 The Author. Published by Sugisaka Masanori at ALife Robotics Corporation Ltd.

This is an open access article distributed under the CC BY-NC 4.0 license

[\(http://creativecommons.org/licenses/by-nc/4.0/\)](http://creativecommons.org/licenses/by-nc/4.0/).

1. Introduction

In Japan, society is aging at an unprecedentedly rapid pace, the average age of the population has been increasing, and these trends are expected to continue. Because of this, researchers have been studying ways of applying information technology (IT) to improving the medical and/or mental support provided to older adults, including persons with extreme psychiatric disorders.

In our previous study [1], we developed a system for analyzing the facial expressions of a person obtained while answering interview questions posed by an animated agent. To accomplish this, we used MikuMikuDanceAgent (MMDAgent) [2], which is a freeware animation program that allows users to create and animate movies with agents. To produce the agent's voice, we used the speech synthesis function setting built into MMDAgent. Agent motions during the interview included "nodding" and "blinking" movements.

In this study, to make the agent's performance on a personal computer (PC) screen more human-like, we have developed a system for agent facial expression generation that uses vowel recognition when generating synthesized speech.

2. Proposed System and Method

2.1. System overview and outline of the method

Figure 1 shows the processing flow of this system, which consists of six processing units:

- creating facial expression data,
- recording voice utterances,
- automatic WAVE file division,
- speech recognition by the Julius high-performance, two-pass large vocabulary continuous speech recognition decoder software [3],
- insertion of expressionless data,
- creating facial expression motion.

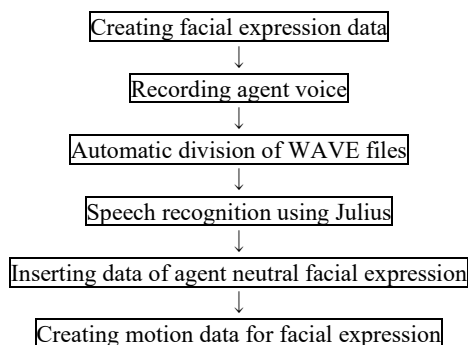


Fig. 1. Processing flow of system.

The facial expression data are created in advance. Other five processing units are successively performed in the order described above.

3. Development of facial expression synthesis system

3.1. Creating facial expression data

Expression motions are generated by combining the expression data of each vowel for each utterance motion. Facial expression data were created with MikuMikuDance [4]. In this study, in order to realize more human-like agent facial expressions, facial expression data are created for the vowels / a /, / i /, / u /, / e /, and / o /, which correspond to all vowels existing in Japanese. Figure 2 shows facial expression of the agent in uttering each vowel.

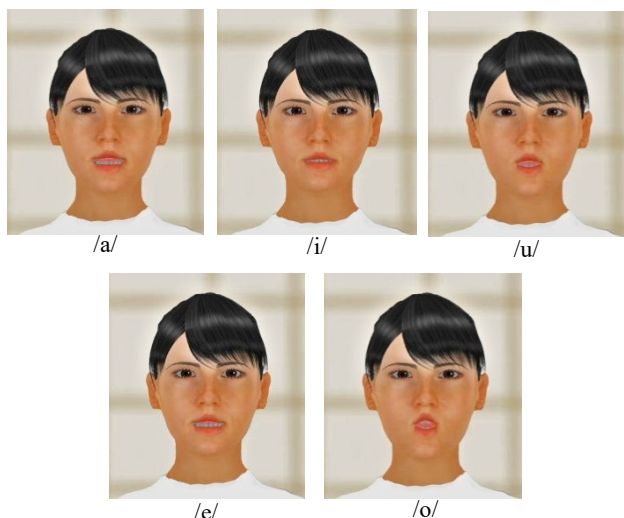


Fig. 2. Facial expression of the agent in uttering each vowel.

3.2. Agent voice recording

In our system, utterance contents are input as text and used by the MMDAgent to output synthesized voice that is then recorded by a stereo mixer inside a PC and saved as a WAVE file.

3.3. Automatic division of WAVE files

After all the utterances have been recorded, if there are multiple questions, the WAVE file is automatically divided for each question, and a new WAVE file is created for each question.

```

sentence: 食欲 (はしかか) ですか。
wseal: <s> 食欲+名詞 (は+助詞 しかか)+形+状詞 です+助動詞 か+助詞 </s>
phseal: sp_S | sh_B o | k | u | y | o | k | u | e | w_B a_E | i_B k_I
cmscore1: T.000 0.938 0.889 0.864 0.986 0.525 1.000
score1: 153.970123 (AM: 314.087799 LM: -160.117676)
=== begin forced alignment ===
-- phoneme alignment --
id: from to n_score unit
-----
0 21 -3.476162 sp_S+sh_B[sp_S]
3 13 2.109969 sp_S-sh_B+o |[sp-sh_B+o_B]
14 19 1.689885 sh_B-o_I+k |[ch_B-o_I+k_B]
20 27 3.027793 o_I-k_I+u |[o_I-k_I+u_B]
28 32 2.053262 k_I-u_I+y |[k_B-u_I+y_B]
33 37 1.453514 u_I-y_I+o |[h_B-y_I+o_B]
38 44 2.031867 y_I-o_I+k |[v_F-o_I+k_B]
45 50 2.800434 o_I-k_I+u |[o_I-k_I+u_B]
51 58 2.530456 k_I-u_E+u_B|[k_B-u_E+u_B]
59 64 2.310862 u_E-w_B+a_E|[h_E-w_B+a_E]
65 73 1.006805 w_B-a_E+i_B|[w_B-a_E+i_B]
74 81 2.154485 a_E-i_B+k |[a_E-i_B+k_B]
82 89 2.765762 i_B-k_I+a |[i_B-k_I+a_B]
90 96 2.810837 k_I-a_I+g |[f_E-a_I+g_E]
97 103 2.868757 a_I-g_I+a |[a_B-g_I+a_B]
104 108 3.200406 g_I-a_E+d_B
109 114 2.010101 a_E-d_B+e |[a_E-d_B+e_B]
115 121 1.140069 d_B-e_I+s_I
122 125 1.529968 e_I-s_I+u |[e_I-s_I+u_B]
126 131 2.791662 s_I-u_E+k |[s_B-u_E+k_B]
132 139 2.428546 u_E-k_B+a |[h_E-k_B+a_E]
140 159 0.035126 k_B-a_E+sp_S |[h_E-k_B+a_E+sp_S]
159 177 0.900673 a_E-sp_S[sp_S]
re-computed AM score: 314.119537
=== end forced alignment ===
  
```

Fig. 3. Results of speech recognition by Julius.

0	o	14
1	o	27
2	o	28
3	u	37
4	o	38
5	o	50
6	u	51
7	u	64
8	a	65
9	a	81
10	i	82
11	i	89
12	a	90
13	a	103
14	a	109
15	a	114
16	e	122
17	e	125
18	u	126
19	u	139
20	a	140
21	a	177

Fig. 4. An output file listing vowels, and their start and end times.

3.4. Speech recognition using Julius

The results of speech recognition using Julius are shown in Figure 3. From the recognition results, which also includes the information of consonants, an output file is created that only lists the vowels, and their start and end times (Figure 4).

3.5. Inserting neutral agent motion data

In order to create more natural agent facial expressions, processing is then performed to insert a neutral facial expression when the vowel / a / is continuous. Figure 5 shows the insertion of neutral motion data executed when “いかがですか?” (“ikaga desu ka?”), which means, “How is it?”, is spoken in Japanese. Since the vowel / a / is continuous when “ka” and “ga” are uttered, neutral data are inserted between them.

3.6. Creating motion data for a facial expression

Figure 6 shows the flow of creating a facial expression motion. The facial expression motion data, in the form of a Vocaloid Motion Data (vmd) file, is created by composing the vowel vmd file based on the speech duration (Sect. 3.4, Figure 4). First, the number of bones of the standard wire frame model to be used is calculated from the number of vowels and neutral frames, and the total number is written in the vmd file header. Next, the vowel voicing time is converted to the number of frames (1 frame = 1/30 seconds) in order to set each bone of the facial expression data and is then written to the vmd file.

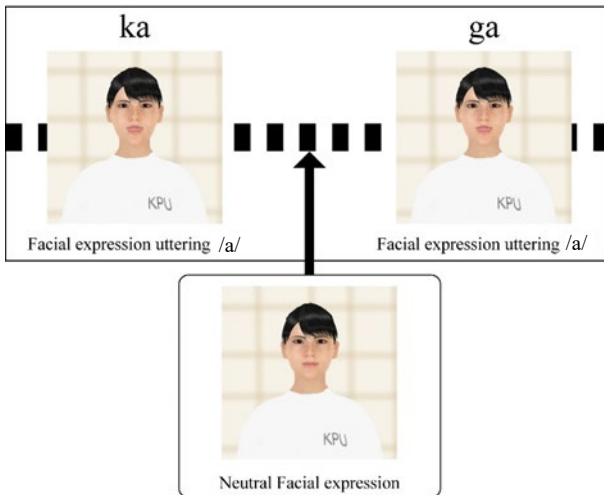


Fig. 5. Insert neutral data while vowel / a / is continuously uttered.

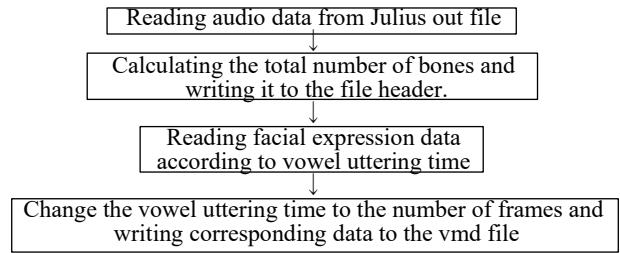


Fig. 6. Flow of facial expression motion creation.

4. Experiment

4.1. Conditions

The experiment was performed on a Dell Inspiron 15 PC equipped with an Intel Core i7-6700HQ 2.2 GHz central processing unit (CPU) and 8.0 GB of random access memory (RAM). The Microsoft Windows 7 Professional operating system (OS) was installed on the PC, and Microsoft Visual C++ 2010 Express was used as the development language.

An animated agent that utters eight questions used in the initial diagnosis of depression by psychiatrists (Table 1) was created under two conditions (Condition 1: Created manually, Condition 2: Created with this system). In addition, an animated agent that asked nine questions (Table 2) on the Hasegawa Dementia Scale [5], which is used in the diagnosis of dementia, was also created using our system (Condition 3).

The content of the questionnaire is “How was the agent’s mouth movements?” and the answer options were from a five-point scale (5: very natural, 4: natural, 3: normal, 2: unnatural, 1: very unnatural).

Table 1. Depression diagnosis test.

No.	Question
1	Do you get depressed or feel gloomy in a daily life?
2	Do you feel less motivated or lacking energy?
3	Do you get enough sleep at night?
4	How is your appetite?
5	Do you enjoy any hobbies on your day off?
6	Do you feel worthless for yourself or hopeless on your alive days?
7	How about concentration and attention to work?
8	Do you work efficiently?

Table 2. Dementia diagnosis test.

No.	Question
1	How old are you?
2	What day of the month, what day, what day is it?
3	Where are we now?
4	Please tell me the three words you will say. Please remember it later.
5	Please subtract by seven in continuous order using 100 as the starting value.
6	Please tell me the numbers I will say from the reverse. 6-8-2.
7	Please say the words that you learned earlier.
8	I will show you five items. Tell me what happened because I'll hide it.
9	Please say as many vegetable names as you know.

4.2. Results and discussion

Table 3 shows the average impression evaluation values of all subjects under Conditions 1 and 2. In Table 3, headings “1” to “8” indicate items by question no., and the “average value” is obtained by averaging over all questions. In the average impression evaluation values shown in Table 3, Condition 2 was higher than Condition 1 for all questions, and the “average value” was about 18% higher. From these results, we confirmed that our system (Condition 2) has an advantage over manual work (Condition 1).

Table 4 shows the average impression evaluation values of all subjects under Condition 3. In Table 4, headings “1” to “9” indicate items by the question no., and the “average value” is obtained by averaging over all questions. From this table, it can be seen that because the animated utterances of the dementia diagnosis questions also obtained high impression evaluation values, the versatility of our system has been demonstrated.

Table 3. Average impression evaluation value of all subjects under Conditions 1 and 2.

Condition	1	2	3	4
1	2.29	2.71	2.86	3.57
2	3.29	3.71	3.29	3.86

Condition	5	6	7	8	Ave.
1	2.71	3.43	2.43	2.57	2.82
2	3.29	3.57	2.57	3.00	3.32

Table 4. Average impression evaluation value of all subjects under Condition 3.

Condition	1	2	3	4	5
3	3.57	3.86	3.57	4.00	3.57

Condition	6	7	8	9	Ave.
3	3.86	3.71	3.86	4.00	3.78

5. Conclusion

Herein, we reported on the development of a system for agent facial expression synthesis generation that uses vowel recognition when generating synthesized speech. The speech is recognized using the Julius high-performance, two-pass large vocabulary continuous speech recognition decoder software system, after which agent facial expression synthesis is performed using preset parameters depending on each vowel sound.

To create the agent, we used MMDAgent, which is a freeware animation program that allows users to create and animate movies with agents. To produce the agent’s voice, we used the speech synthesis function setting built into MMDAgent. The impression evaluation values obtained from a questionnaire survey indicate that an agent produced by our proposed system is more natural than an agent created using preset parameters manually decided for each utterance. In the future, we plan to use this system for facial expression analysis and speech analysis experiments. It is believed that this system could prove more useful in the treatment of persons suffering from psychiatric ailments such as depressive disorder and dementia than our reported system.¹

Acknowledgments

The authors would like to thank Professor J. Narumoto of the Kyoto Prefectural University of Medicine for his valuable support and helpful advice during the course of this research. We would also like to thank the subjects of our experiments for their cooperation. This research was supported by the Center of Innovation Science and Technology based Radical innovation and Entrepreneurship Program (COI STREAM) of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. T. Asada, D. Kogi, R. Shimada, Y. Yoshitomi, and M. Tabuse, A System for Analyzing Facial Expression and Verbal Response of a Person while Answering Interview Questions by Agent, in *Proc. of Int. Conf. on Artif. Life and Robotics* (Beppu, Oita, Japan, 2018), 534-537.
2. MMDAgent, <http://www.mmdagent.jp/> Accessed 6 July 2019.

3. Julius, <http://Julius.osdn.jp/>, Accessed 6 July 2019.
4. MikuMikuDance, <https://sites.google.com/view/vpvp/>, Accessed 9 December 2019.
5. Chiba medical association Revised Hasegawa's dementia scale (HDS-R) https://www.jpn-geriat-soc.or.jp/tool/pdf/tool_05.pdf, Accessed 6 July 2019.

Authors Introduction

Dr. Taro Asada



He received his B.S., M.S. and Ph.D. 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. HIS, IEEEJ member

Ms. Ruka Adachi



She received her B.S. degree from Kyoto Prefectural University in 2018. She works at Software Service, Inc.

Mr. Syuhei Takada



He received his B.S. degree from Kyoto Prefectural University in 2019. He works at Seika Town Hall.

Dr. Yasunari Yoshitomi



He received his B.E, M.E. and Ph.D. 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. IEEE, HIS, ORSJ, IPSJ, IEICE, SSJ, JMTA and IEEEJ member.

Dr. Masayoshi Tabuse



He received his M.S. and Ph.D. 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. IPSJ, IEICE and RSJ member.