Music Recommendation System Driven by Interaction between User and Personified Agent Using Speech Recognition, Synthesized Voice and Facial Expression

Ayumi Matsui[1], Miki Sakurai[2], Taro Asada[3], Yasunari Yoshitomi[3], Masayoshi Tabuse[3]
1: Sumitomo Mitsui Card Co., Ltd., 4-5-15 Imabashi, Chuo-ku, Osaka 541-0042, Japan
2: TIS Inc.,8-17-1 Nishi-Shinjuku, Shinjuku-ku, Tokyo 160-0023, Japan,
3: Graduate School of Life and Environmental Sciences, Kyoto Prefectural University,
1-5 Nakaragi-cho, Shimogamo, Sakyo-ku, Kyoto 606-8522, Japan
E-mail: t_asada@mei.kpu.ac.jp, {yoshitomi, tabuse}@kpu.ac.jp
http://www2.kpu.ac.jp/ningen/infsys/English_index.html

Abstract

We propose a music recommendation system that is characterized by an interaction between a user and a personified agent. Speech is recognized using a speech recognition system called Julius, and the facial expression of the agent is then synthesized using preset parameters that depend on each vowel. We add a new function that changes the facial expression of the agent according to the response of the user to music recommended by our previously proposed system. The effectiveness of the proposed system is verified.

Keywords: Music recommendation, MMDAgent, Facial expression synthesis, Speech recognition, Speech synthesis.

1. Introduction

In Japan, the average age of the population has been increasing, and this trend is expected to continue. This trend is more pronounced in rural areas. Recently, music therapy has been used to improve the recognition ability of people, particularly older people. Music therapy may be more effective if music that is liked by an individual is adopted. We have previously developed a music recommendation system that aims to improve recognition ability, including a system through the Internet using a videophone system.[1-3] In our previously proposed system,[1-3] it is necessary to input the subjective evaluation of the user into a computer to determine the next music that is to be recommended. User input is inconvenient for elderly people, especially those with dementia.

In the present study, to overcome this inconvenience, we propose a system for music recommendation that is characterized by an interaction between a user and a personified agent, which uses speech recognition, synthesized voice and facial expression generation.

2. Proposed System and Method

2.1. *System overview*

Figure 1 shows the processing flow of the proposed system, the music recommendation module of which is based on the previously proposed system[2] that used collaborative filtering and impression words (see the paper[2] for detail on the music recommendation module). The proposed system is characterized by an interaction between a user and a personified agent, and uses speech recognition, synthesized voice and facial expression generation for integrative music recommendation.

To make the agent on a personal computer (PC) screen appear more human-like, we developed a system for agent facial expression generation that uses vowel recognition when generating synthesized speech.[4] We added a new function to change the facial expression of the agent according to the response of the user to music recommended by our system.

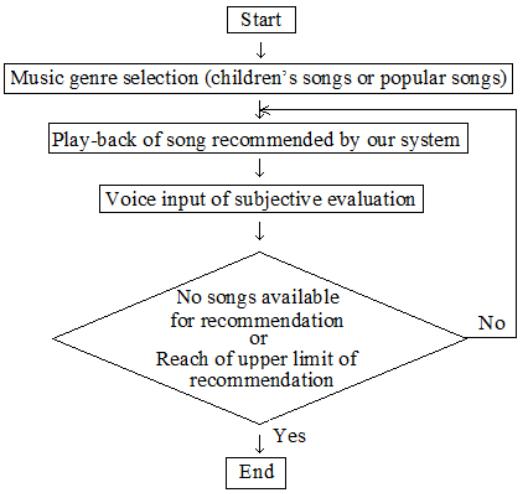

Fig. 1. Processing flow of the proposed system.

2.2. *Personified agent*

Figure 2 shows the process for agent generation in our system, which consists of six steps: creating facial expression data, recording vocal utterances, automatic WAVE file division, speech recognition by the Julius[5], insertion of expressionless data, and the creation of facial expression motion.[4] The facial expression data are created in advance.

Expressive motions are generated by combining the expression data of each vowel for each utterance motion. Then, utterance contents are input as text and used by the MikuMikuDanceAgent (MMDAgent),[6] which is a freeware animation program that allows users to create and animate movies with agents, to output synthesized voice that is then recorded by a stereo mixer inside a PC and saved as a WAVE file. Speech is recognized using a speech recognition system called Julius,[5] followed by facial expression synthesis of the agent using preset parameters depending on each vowel. Facial expression data were created with MikuMikuDance.[7] In this study, in order to generate more human-like agent facial expressions, facial expression data were created for the vowels / a /, / i /, / u /, / e /, and / o / (Fig. 3).[4] In order to create more natural agent facial expressions, processing is then performed to insert a neutral facial expression when the same vowel, for example / a /, is continuous.[4]

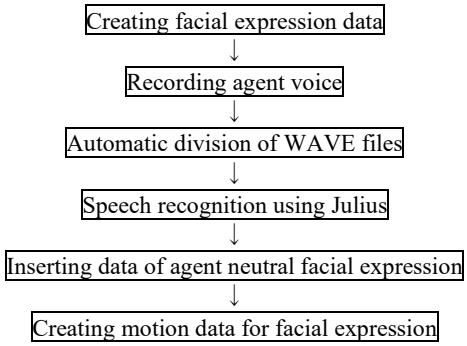

Fig. 2. Processing flow of agent generation in our system.[4]

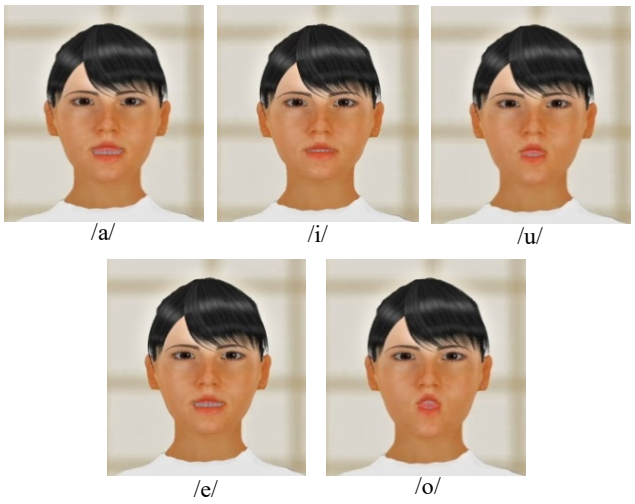

Fig. 3. Facial expression of the agent in uttering each vowel.[4]

Figure 4 shows the flow of creating a facial expression motion (see the paper[4] to understand the processing described in Fig. 4 in more detail).

2.3. *Navigations by improved agent*

In the music-recommendation process, all user navigations are performed by the synthetic voice of the agent, appearing on the PC screen facing the user. All dialogue spoken by the agent is situationally selected by the proposed system. The user's answers to the questions generated by the agent are recognized using the voice recognition function of the system, and the agent motions, including facial expressions, are then generated.

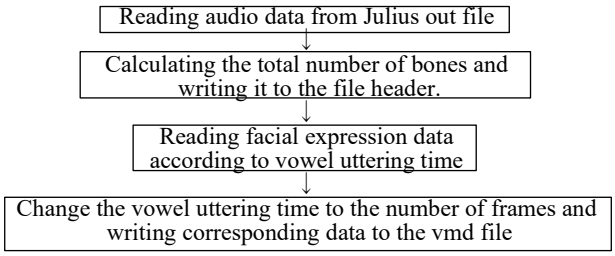

Fig. 4. Flow of facial expression motion creation.[4]

(a) (b)

Fig. 5. Snapshots in the respective reactions of the agent after recognizing: (a) a positive answer, and (b) a negative answer.

Figure 5 shows two snapshots of the reaction of the agent after recognizing: (a) a positive answer, i.e., the user wishes to listen to the recommended song again in the future, and (b) a negative answer, i.e., the user does not wish to listen to the recommended song in the future. In the case of (a), the agent nods twice and raises the corners of the mouth slightly, while in the case of (b), the agent also nods twice, but instead lowers the corners of the mouth slightly. Figure 6 shows a snapshot of music recommendation being performed by the proposed system.

Fig. 6. Snapshot in performing song-recommendation by the proposed system.

3. Experiment

3.1. *Conditions*

Since older people tend to prefer children's songs,[8] we selected a CD[9] described as an anthology of older songs that are enjoyed by older people with dementia, and selected (1) 52 songs on the CD that were also included in a musical textbook database[10] for elementary schools, and (2) 58 other popular songs on the CD.

The experiment was performed on a Dell Inspiron 15 PC, equipped with an Intel Core i7-6700HQ 2.60 GHz central processing unit (CPU) and 8.0 GB of random access memory (RAM). The Microsoft Visual C++ 2010 Express and Visual C++ 6.0 were used as the development languages.

To evaluate the proposed method described in Section 2, 10 males (Subjects A to G in their 20s, I in his 40s, J in his 50s, and L in his 80s) and 2 females (Subject H in her 20s and K in her 50s) participated in the experiments. Each of the subjects navigated the music-recommendation process of the agent. The upper limit for the number of songs recommended by the proposed system was set to be 15. The experiment was performed separately using the database consisting of the 52 children's songs, and that consisting of the 58 popular songs, with two experimental conditions for each (Condition 1: with the input to the system being performed by a supporter instead of the agent navigating the system, and Condition 2: with the agent navigating the proposed system). Following the experiment, all subjects were requested to select one of five [evaluation value] answers ([5] absolutely yes, [4] yes, [3] I can't say either, [2] no, [1] absolutely no) to seven questions (Table 1) in order to evaluate the proposed system. The subjects were also asked to provide comments on the system.

Table 1. Questionnaire to evaluate the proposed system.

No.	Question
1	Was music-recommendation on condition 2 smoother than that on condition 1?
2	Were explanations by the agent easy to understand?
3	Were dialogues with the agent natural?
4	Were movements of agent mouth natural?
5	Were agent's reactions natural after recognizing user's positive answer for listening to the just recommended music again in the future?
6	Were agent's reactions natural after recognizing user's negative answer of no more he just recommended music from now on?
7	Did you feel enjoyable in using the proposed system?

3.2. *Results and discussion*

Table 2 shows the average-values for each question listed in Table 1. The mean value of the averages listed in Table 2 was 4.1, suggesting a positive overall evaluation of the proposed system. For questions 5 and 7, the evaluation averages were very large, while those of questions 1 and 4 were lower.

Table 2. Evaluation of the proposed system.

Question no.	1	2	3	4	5	6	7
Average	3.6	4.4	4.0	3.6	4.5	4.3	4.5

Table 3 shows the average values for each gender (M=male, F=female) and age group for each of the questions listed in Table 1. For Male Subject L, who was in his 80s, the evaluation values for questions 1 and 2 were very low compared with those of other subjects. Subject L described in his comments that the agent spoke too rapidly for him to understand its statements, and that it was difficult to input his voice using the microphone. Three male subjects (D and E in their 20s and Subject J in his 50s) also described in their comments that it was difficult to input their voices using the microphone. Consequently, speech recognition for Subjects D, J, and L was not performed smoothly by the proposed system, resulting in inputting voice again after training the system. Nearly all of the comments by the subjects expressed positive evaluation of the proposed system, except in relation to voice input to the system.

Table 3. Evaluation of the proposed system by each age group.

Question no.		1	2	3	4	5	6	7
Average	M-20s (n=7)	3.7	4.7	4.4	3.6	4.6	4.3	4.1
	F-20s (n=1)	4	5	3	4	5	5	4
	M-40s (n=1)	3	4	4	3	4	4	4
	M-50s (n=1)	4	4	4	4	4	4	5
	F-50s (n=1)	5	5	5	5	5	5	5
	M-80s (n=1)	2	2	4	4	4	4	5

4. Conclusions

We proposed a system for music recommendation, characterized by an interaction between a user and a personified agent that uses speech recognition, synthesized voice and facial expression generation. Speech is recognized using a speech recognition system called Julius, and the facial expression of the agent is then synthesized using preset parameters that depend on each vowel. We used MMDAgent to create the agent. To produce the voice of the agent, we used the built-in speech synthesis function setting in MMDAgent. We added a new function that changes the facial expression of the agent according to the responses of the users to music recommended by our previously proposed system. The effectiveness of the proposed system was verified.

Acknowledgements

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 COI STREAM of the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. C. Koro, Y. Yoshitomi, T. Asada, and S. Yoshizaki, Music recommendation aimed at improving recognition ability using collaborative filtering and impression words, *in Proc. 17th Int. Symp. on Artificial Life and Robotics*, ed. M. Sugisaka (Japan, Beppu, 2012), pp. 222–225.
2. S. Yoshizaki, Y. Yoshitomi, C. Koro, and T. Asada, Music recommendation hybrid system for improving recognition ability using collaborative filtering and impression words, *J. Artif. Life and Robotics* **18**(1-2) (2013) 109–116.
3. Y.Yoshitomi, T.Asada, R.Kato, Y.Yoshimitsu, M.Tabuse, N.Kuwahara, and J.Narumoto, Music recommendation system through Internet for improving recognition ability using collaborative filtering and impression words, *J. Robotics, Networking and Artif. Life*, **2**(1), 2015, 54–59.
4. T. Asada, R. Adachi, S. Takada, Y. Yoshitomi, and M. Tabuse, Facial expression synthesis using vowel recognition for synthesized speech, *in Proc. 2020 Int. Conf. on Artificial Life and Robotics*, ed. M. Sugisaka (Japan, Beppu, 2020), pp.398‑402.
5. Julius, http://Julius.osdn.jp/, Accessed 24 November 2020.
6. MMDAgent, http://www.mmdagent.jp/ Accessed 24 November 2020.
7. MikuMikuDance, https://sites.google.com/view/vpvp/, Accessed 29 November 2020.
8. T. Takahashi, Research report on songs familiar to people advanced in years (in Japanese), *J. Japanese music therapy associate* **15**(1) (1997) 68–75.
9. T. Akahoshi, *Good old anthology enjoyable for people advanced in years and troubled with dementia* (in Japanese) (Kirara shobo, Tokyo, 2009).
10. Music textbook database for elementary school by Kanagawa prefectural education center (in Japanese), http://kjd.edu-ctr.pref.kanagawa.jp/daizai_music/, Accessed 27 November 2020.