Possibility of feelings expression in anthropomorphic agent by synthetic sound

Mayumi Koga*, Kazuyuki Miura*, Atsushi Osa*, Hidetoshi Miike*

*Graduate school of Science and Engineering, Yamaguchi University, 2-16-11, Tokiwadi, Ube-shi, Yamaguchi, 755-8611, JAPAN j006vm@yamagichi-u.ac.jp

Abstract: In recent years, the progress of computer technology has brought the incidence of digital divide. The development of "anthropomorphic interface" is enumerated as one of the solutions to relieve from this problem. The anthropomorphic agents are expected to have expression of feelings like humans. We have focused on voices of exclamations. In a previous study, we analyzed prosodic features of several exclamation voices, and generated synthesize sounds according to the extracted features. As the results, the voices were confused to evoke "happy" emotion with "surprise" emotion easily. The synthetic sounds evoked "happy" emotion mainly. Thus we found a possibility of feeling expression by sound. In this study, we focused the following 2 topics: (1) changing ambiguity of expressed emotions, (2) the best timing of combination an agent's voice/sound and body motion. Then, we discuss about possibility of the synthetic sound in the anthropomorphic interface. A result of the exclamation voice showed a similar tendency that the evaluated values were decreasing according to the time lag from the synchronization. To the contrary, in the synthetic sounds, the tendency was difference from the result of the exclamation voice. The best timing in the synthetic sound and the exclamation voice were perhaps different. Moreover, we investigated an interaction between the ambiguous voices and motions of an anthropomorphic agent to pursue the difference of the exclamation voice and the synthetic sound. However, the synthetic sound with the animation expressed more ambiguous feelings than the exclamation voice with the animation.

Key words: *Anthropomorphic interface, Feelings expression, Exclamation voice, Synthesize sound, Timing*

1. Introduction

Recently, there are various information equipments around us along with computer technology. These equipments have brought convenience for us. To the contrary, it is also a fact that the equipments become excessively complex, and the incidence of digital divide is feared. The development of "anthropomorphic interface" is enumerated as one of the solutions to relieve from this problem [1]. The anthropomorphic agents are expected to provide proper functions and information adjusting user's skill on these equipments, and have expression of feelings like humans [2]. The agent makes a sense of affinity with the equipments; moreover it would resolve psychological stress for using the equipments. In a previous study, we have focused on a feeling expression of the agent, above all, expressions of exclamation [3]. The feeling expression using exclamation voices depends not only on these meanings but also on these prosodic features. The prosodic features indicate affective information in voices [4][5]. We analyzed the prosodic features of several exclamations, and generated synthesize sounds according to the extracted features. As the results, we found that some synthetic sounds did not evoke the same emotion as their original exclamation voices did. The original exclamations were confused to evoke "happy" emotion with "surprise" emotion easily, but the synthetic sounds evoked "happy" emotion mainly [3]. So in the exclamations, we thought that language/vocal sound evoked ambiguous impression. And in the synthetic sounds, we found possibility that the synthetic sound can evoke a clear emotion, because the sound doesn't have any meaning as language. In addition to that, the synthetic sound has some special characteristic: the sound don't depend on sex, age, type etc.

On the other hand, we thought that combination of animations of an anthropomorphic agent and the voices might influence the evoked emotion. It is well known that, in the human sense, visual perception occupies about 80% and the rest is mostly auditory perception [6]. And body motion, that is a nonverbal communication, occupies approximately half of the human's communications transmission [7]. And, we interested in the best timing of the voices and the animations.

Temporal structure of visual and auditory stimulation is important for sensory integration of visual and auditory perception [8].

In this study, our purpose was to find difference between voices and these synthetic sounds under combination of an agent's voice/sound. The following 2 topics were especially focused:

(1) changing ambiguity of expressed emotions,

(2) the best timing of combination an agent's voice/sound and body motion.

Then, we discuss about possibility of the synthetic sound in the anthropomorphic interface.

2. Evaluation experiment I

(About expression of "sadness")

2.1 Method

We researched relationships between timing and "sadness", timing and "hedonic scale".

We selected an exclamation voice that was identified only "sadness" and generated a synthetic sound based on the prosodic features of the exclamation voice. And we selected a character who didn't have facial expression and androgynous childe because that we wanted to add body motions and voices sounds was important. We prepared 16 animations, in which the character looked at the ground at a slow speed because the motion obviously gave the impression of "sadness" to observers (see Fig 1). The character said the exclamation voice or the synthetic sound. Each animation was arranged to have different timing between its movement and its voice/sound, from synchronization to 240 frames difference, 30 frames interval. 10 observers watched these animations, and they rated expression of the character according to "sadness" level, and "hedonic scale" with 7 points evaluation. The observer was instructed in a situation that the observers started a conversation with the character.

2.2 Results and Discussions

Figure 2,3 shows the evaluated value of "sadness", and "hedonic scale" about both the exclamation voice and the synthetic sound in each of Timing. 2-factors ANOVA and Tukey's HSD was used. As a result, there was main effect about timings in "sadness" [$F(7,63)=4.672$, $p<.05$]. And, there was main effect of voice/sound in "sadness" [$F(1,9)=39.041$, $p<.05$]. As a result of Tukey's HSD [HSD=1.195], "sadness" of the sound in several timings (Timings 1,3,4,8) were significantly lower than these of the voice in the same timings. However, the tendency of the best timing seemed to have special timings at Timing 2, and around Timing5. This tendency perhaps indicates an interesting characteristic of the sound.

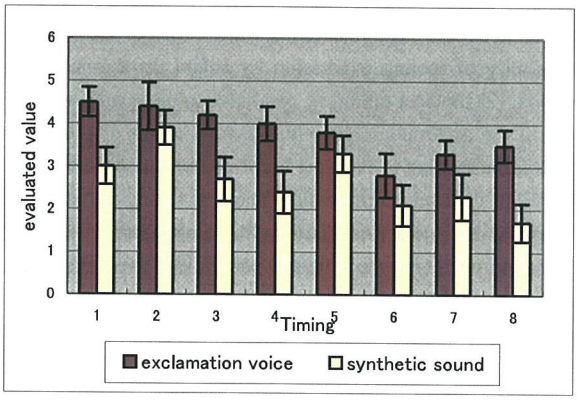

Figure 2 : Evaluation result of "sadness"

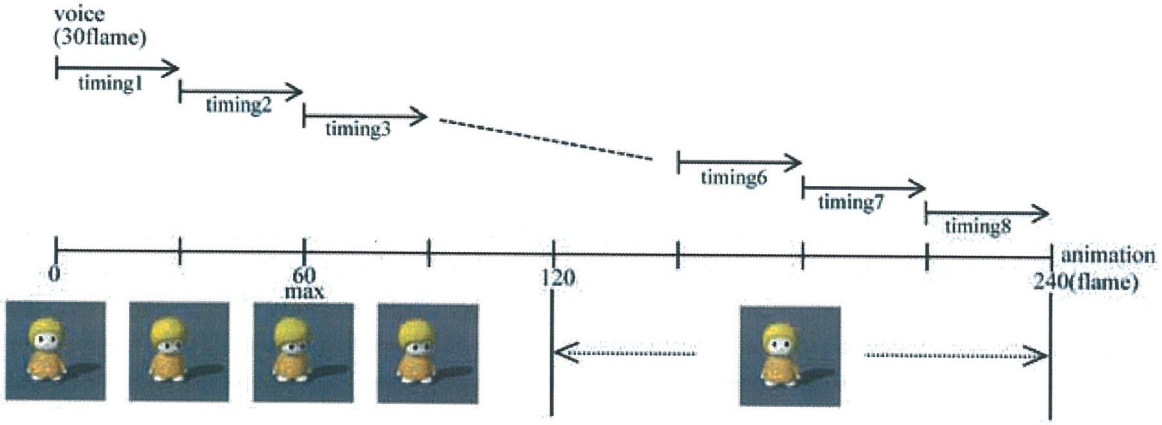

Figure 1: Character animation for experiment I

In "hedonic scale", there was no main effect of timings.

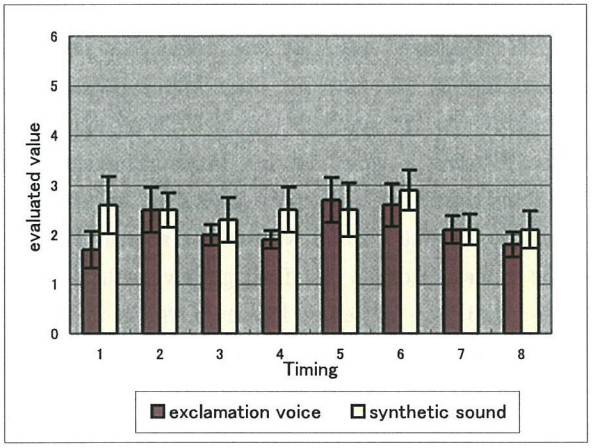

Figure 3: Evaluation of "hedonic scale"

3. Evaluation experiment Ⅱ

(About expression of "happy" and "surprise")

3.1 Method

We selected an exclamation voice that was confused "happy" with "surprise" in the result of our previous study [3]. A synthetic sound was generated using the prosodic features based on this exclamation voice. And we selected the same character of experiment Ⅰ. We prepared 16 animations, in which the character reached up like joy when the right arm was raised highest to evoke the impression of "happy" and/or "surprise" (see Figure 4). And the character produced the exclamation voice or the synthetic sound. Each animation was arranged to have different timing between its movement and its voice/sound, from the synchronization to 90 frames difference, with 10 frames interval. Character's motion in the animation of experiment Ⅰ was three times that of this experiment (experiment Ⅰ : 120 frames, experiment Ⅱ : 40 frames). 10 observers watched these animations, and they rated expression of the character according to "happy" level, "surprise" level, and "hedonic scale" with 7 points evaluation. In the experimental instruction we directed the observers to image a situation that they started a conversation with the character.

3.2 Results and Discussions

(1) Difference of impressions about "happy", "surprise", and "hedonic scale" by shifts timings

Figure 5 shows the evaluated value of "happy" about both the exclamation voice and the synthetic sound for each of timing. 2-factor ANOVA reported measures, and Tukey's HSD were used. As a result, there was the no main effect of timings in "happy" [$F(7,8)=3.300$, $p<.05$]. In the results of the exclamation voice, the evaluated value of "happy" showed the tendency to decrease (Fig 5) according to the time lag. But, as a result of Tukey's HSD [HSD=1.088], Timing 8 in the exclamation voice that have the lowest evaluated value differed significantly from Timings 1~4, 6, Timing 2 in the exclamation voice that have the highest evaluated value differed significantly from Timings 5, 8. There was no main effect of voice/sound [$F(1,9)=2.283$, $p<.05$], and there was no interaction between Timings and voice/sound [$F(7,63)=1.655$, $p<.05$]. However, the evaluated values of the synthetic sound in Timings 1, and 2 seemed lower than the evaluated values of the exclamation voices in the same timings. The tendency of the best timing perhaps is different between the voice and the sound.

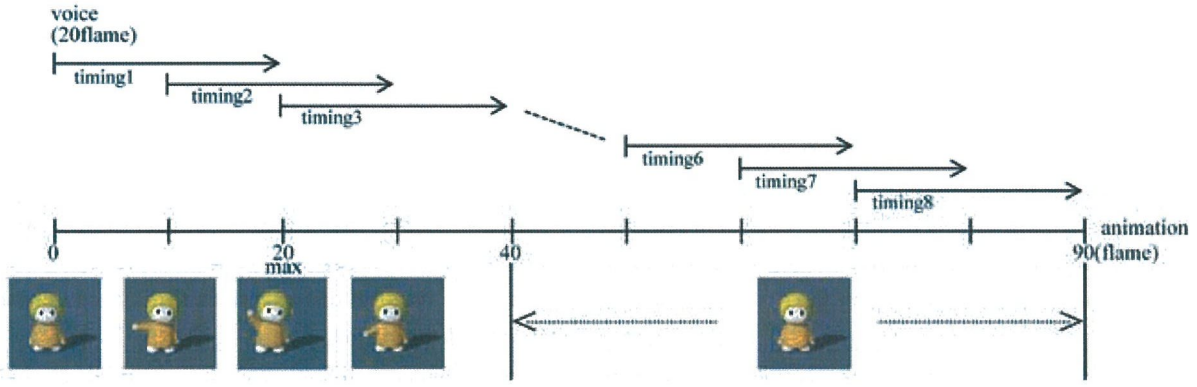

Figure 4: Character animation for experiment Ⅰ

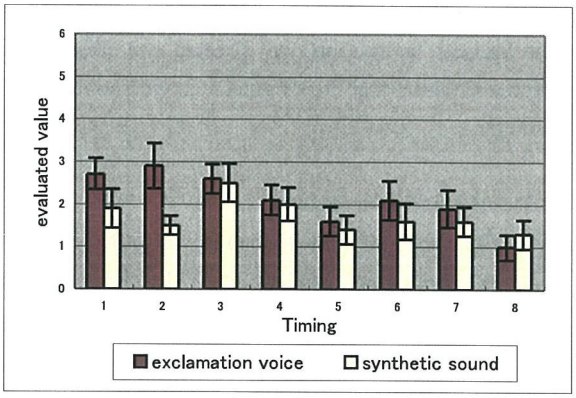

Figure 5 : Evaluation result of "happy"

Figure 6 shows the evaluated value of "surprise" of both the exclamation voice and the synthetic sound in each of Timing. As a result, there was main effect of timing [$F(7,63)=2.194$, $p<.05$]. And as a result of Tukey's HSD [HSD=.9822], Timing 8 in the exclamation voice that have lowest evaluated value differed significantly from Timings 1~3, 6, 7. Timing 2 in the exclamation that have the highest evaluated value differed significantly from Timings 5, 8. To the contrary, every combination of Timing in the synthetic sound showed no significant differences. There was main effect of voice/sound [$F(1,9)=16.148$, $p<.05$], and there was no interaction between factors [$F(7,63)=2.194$, $p<.05$]. The evaluated values of "surprise" in the synthetic sounds were lower than these in the exclamation voices.

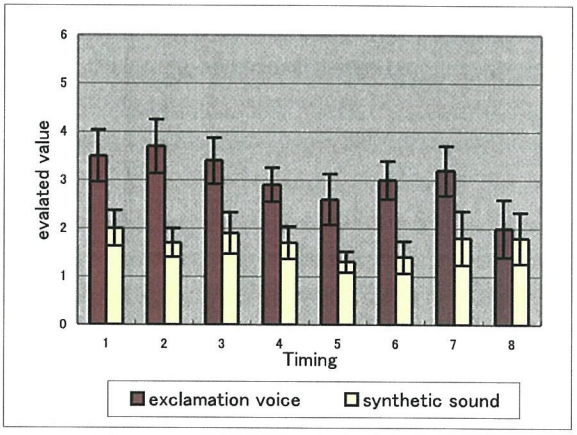

Figure 6 : Evaluation result of "surprise"

Figure 7 shows the evaluated value of "hedonic scale" about both the exclamation voice and the synthetic sound in each of Timing. As the result, there was main effect of timings [$F(7,63)=7.118$, $p<.05$]. There was no main effect of voice/sound [$F(1,9)=.072$, $p<.05$], and there was no interaction between Timings and voice/sound [$F(7,63)=1.696$, $p<.05$]. And, as a result of Tukey's HSD [HSD=.7282], Timing 8 in the exclamation voice differed significantly from Timings 1~3. Timing 2 in the exclamation voice differed significantly from Timings 5,7,8. "Hedonic scale" of the voice was decreasing according to the time lag from the synchronization. Timing 6 in the synthetic sound differed significantly from Timing 3, and Timing3 in the synthetic sound differed significantly from Timing 2,4~8. So, we thought that the synthetic sound was comfortable in Timing 3. The difference of the best Timing between the voice and the sound were not clear. However, "hedonic scale" of the voice and the sound were especially difference in Timing 2.

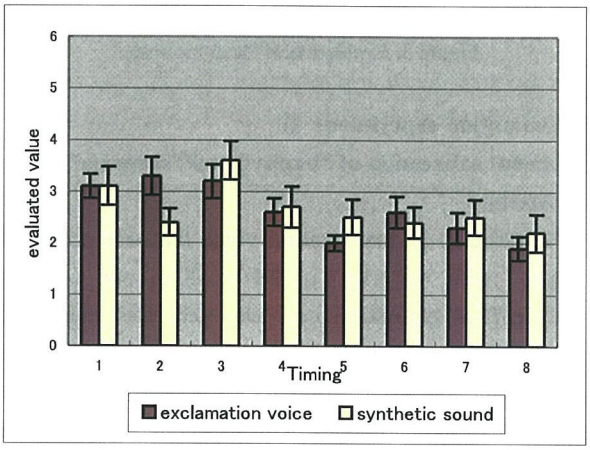

Figure 7 : Evaluation result of "hedonic scale"

(2) Influence on feeling expression by adding animation

2-factor reported measures ANOVA was used. As a result, there was main effect of emotion ("happy" and "surprise") in the exclamation voice [$F(1,14)=10.302$, $p<.05$]. Table 1 shows the result of Tukey's HSD [HSD=0.652], in which each value means the result of subtracted HSD from the absolute value of difference between an average of "happy" and an average of "surprise" the exclamation voice. And, if the result is positive, it meant an existence of the significant difference among these timings. Table 1 showed that there was a significant difference between "happy" and "surprise" in each of the same Timings (see colored values in Table 1). The exclamation voice with animation was identified as emotion of "surprise". And according to the early study [9], we were a comfortable working speed, so we thought that the motion speed in the animation used by this experiment influenced observers feeling identification.

To the contrary, there was not main effect of "happy"

and "surprise" emotions in the synthetic sound. The exclamations voice without animation was confused to evoke "happy" emotion with "surprise" emotion easily, and the synthetic sounds without animation evoked "happy" emotion mainly [3]. The results of this experiment were opposite and interesting. The character of animation was a pretty child, and the synthetic sound was very far from the image of the character. This mismatch might affect the impression of the sound.

Table 1: Difference of between "happy" and "surprise" in exclamation voice

		happy							
	Timing	1	2	3	4	5	6	7	8
surprise	1	.15*	-.42	-.63	.75*	1.25*	.75*	.95*	1.85*
	2	.35*	.15*	-.53	.95*	1.45*	.95*	1.15*	2.05*
	3	.05*	-.15	.15*	.65*	1.15*	.65*	.85*	1.75*
	4	-.45	-.65	-.35	.15*	.65*	.15*	.35*	1.25*
	5	-.55	-.35	-.65	-.15	.35*	-.15	.05*	.95*
	6	-.35	-.55	-.25	.25*	.75*	.25*	.45*	1.35*
	7	-.15	-.05	-.05	.45*	.95*	.45*	.65*	1.55*
	8	.05*	-.05	-.05	-.55	-.25	-.55	-.55	.35*

*$p<.05$

4. Conclusions

In this study, we focused on feeling expressions by body motion and voice/sound of a character. And, we investigated feeling expression in the synthetic sound for finding possibility of feeling expression without voice and language. The results in the exclamation voice showed a similar tendency that the evaluated values were decreasing according to the time lag from the synchronization. The results in the synthetic sound, the tendency were difference from the result of the exclamation voice. The best timing in the synthetic sound and the exclamation voice were perhaps different. On the other hand, the synthetic sound with the animation expressed more ambiguous feelings than the exclamation voice with the animation. This results were opposite from the results of our previous study in that we compared the synthetic sound and the exclamation voice without animation. We assumed that the animation could restrict the feeling clearly. However, the mismatch of the character and the synthetic sound might affect the feeling impression of the observers. These results showed that it is difficult to use the synthetic sound as a substitute of voice. We think that the synthetic sound can used as sound effects of motion to express feelings of characters.

In future work, we have to discuss possibility the synthetic sound as sound effects including emotions. Character's motion speed and pitch/length of the synthetic sound should be researched for manipulation of the agent's feeling expression.

5. References

[1] Hiroshi Dohi, Mitsuru Ishizuka: Human-Agent Interaction with a Life-like Character Linked with WWW: Japan Society for Artificial Intelligence, Vol.17 No.6, pp.693-700 (2002) (in Japanese)

[2] Yousuke Hiruma, Yoshihiro Adachi, Shigeo Morisita: An evaluation of multimodal impression by presenting an emotional voice and an expression face simultaneously: TECHNICAL REPORT OF IEICE. HCS2004-23, pp.7-12 (2004) (in Japanese)

[3] Mayumi Koga, Kazuyuki Miura, Atsushi Osa, Hidetoshi Miike: Analysis of features and possibility of speech synthesis for feelings speech by exclamation: Proceedings of the 8th Annual Conference of JSKE2006, p.85 (2006) (in Japanese)

[4] Shuichi Itabashi: Sound Engineering. MORIKITA Publishing Co., Ltd.; p.58 (2005) (in Japanese)

[5] Shoichi Takeda, Hidefumi Yamato, Nao Kaneko, Kazuaki Yamamoto, Teruo Muraoka, Muhd Dzulkhiflee Hamzah: Prosodic Feature of *Kyogen* Speech with "Anger", "Joy", and "Sadness" Compared accoding to the Degree of Emotion: Proceeding of The Acoustical Society of Japan2005springtime, pp.209-210 (2005) (in Japanese)

[6] Shin Hasegawa: Image Engineering. CORONA Publishing Co., Ltd.; (1991) (in Japanese)

[7] Takao Kurokawa: Nonverbal Interface. Ohmsha Publishing Co., Ltd.; (1994) (in Japanese)

[8] G.J. Thomas, D. Sanabria, S. Soto-Faraco: Experimental study of the influence of vision on sound localization: J.Exp. Psychol. 28, pp163-177 (1941)

[9] Kumi Naruse: Effect of Different Movement Paces on Mood States: Japan Society of Physical Anthropology, Vol.10, No.4, pp.25-32 (2005) (in Japanese)