Hindawi Publishing Corporation
International Journal of Computer Games Technology
Volume 2008, Article ID 281959, 9 pages
doi:10.1155/2008/281959

Research Article

Generation of Variations on Theme Music Based on Impressions of Story Scenes Considering Human's Feeling of Music and Stories

Kenkichi Ishizuka and Takehisa Onisawa

Graduate School of Systems and Information Engineering, University of Tsukuba, Tennodai 1-1-1, Tsukuba 305-8573, Japan

Correspondence should be addressed to Takehisa Onisawa, onisawa@iit.tsukuba.ac.jp

Received 31 July 2007; Accepted 17 October 2007

Recommended by Kevin Kok Wai Wong

This paper describes a system which generates variations on theme music fitting to story scenes represented by texts and/or pictures. Inputs to the present system are original theme music and numerical information on given story scenes. The present system varies melodies, tempos, tones, tonalities, and accompaniments of given theme music based on impressions of story scenes. Genetic algorithms (GAs) using modular neural network (MNN) models as fitness functions are applied to music generation in order to reflect user's feeling of music and stories. The present system adjusts MNN models for each user on line. This paper also describes the evaluation experiments to confirm whether the generated variations on theme music reflect impressions of story scenes appropriately or not.

1. INTRODUCTION

Music, pictures, and/or text information are combined into multimedia content with interaction among them [1]. The effectiveness of multimodal communication using combined different modal media has been analyzed in the field of cognitive psychology [1]. It is expected that multimodal communication will be performed in everyday life in the future owing to the development of information technology [2]. However, the interaction among different modal media is not necessarily generated by their simple and random combination. Features and impressions of media should be considered well in order to create effective multimedia contents. Therefore, creation of multimodal contents costs more time and labor than that of single-modal one. Support systems for creation of multimodal contents or for the flexible combination of different modal media are taken interest in [3, 4].

The authors are studying on the construction of a system which generates variations on theme music fitting to each story scene represented by texts and/or pictures [5]. This system varies melodies, tempos, tones, tonalities, and accompaniments of a given theme music based on impressions of story scenes. This system has two sections representing (a) relations between story scenes and musical images and (b) relations between features of variations and musical impressions. Since human feeling of stories and music is different among people [6] and the difference is important in multimedia content creation, it is necessary to consider the above relations depending on each user. Although in [5] these relations are obtained by questionnaire data, that is, off line, in the present paper a method, which adjusts the relations for each user on line, is proposed. In this paper, the transformation of theme music is defined as follows. Tunes, tones, musical performances, rhythms, tempos are varied according to story scenes [7].

2. OUTLINE OF PRESENT SYSTEM

2.1. Inputs and outputs

Inputs to the present system are original theme music and numerical information on given story scenes. Outputs are MIDI files of variations on original theme music generated according to each story scene. This paper deals with generation of variations on theme music fitting to stories obtained by the system [8] that generates story-like linguistic

Table 1: Information on story scene.

No.	Information	Num	Information
1	Happiness	7	Kind of character 1
2	Sadness	8	Kind of character 2
3	Surprise	9	Impressions of behavior
4	Fear	10	Kind of character 1 (previous)
5	Anger	11	Kind of character 2 (previous)
6	Disgust	12	Impressions of behavior
—	—	13	Picture's sequence

expressions given four pictures. In this paper, a scene is defined as each picture for story generation. Information on a picture scene, for example, character's emotion, kinds of characters, impressions of character's behavior in a story scene (e.g., violent behavior), picture sequence, as shown in Table 1, is acquired from each picture [8]. These are inputs to the present system.

2.2. System structure

The present system consists of two sections, a musical image acquisition (MIA) section and a theme music transformation (TMT) section as shown in Figure 1. The MIA section converts information on story scenes shown in Table 1 into transformation image parameters (TIPs) by modular neural network (MNN) models [9]. The TMT section transforms inputted original theme music based on values of TIPs, and generates a set of midiformatted candidates of variations on theme music for each story scene. The TMT section applies genetic algorithms (GAs) to the generation of variations candidates, which has MNN models as fitness functions. MNN models consist of three neural network models, an average model network (AMN), an individual variation model network (IVMN), and gating networks. AMN is a hierarchical neural network model expressing user's average feeling of music and stories. IVMN is a radial basis function network model expressing differences among users' feeling of music and stories. The gating network switches over between AMN and IVMN. The present system adjusts IVMNs and the gating networks for each user.

3. MUSICAL IMAGE ACQUISITION (MIA) SECTION

The MIA section is constructed by MNN models. The inputs to MNN models are shown in Table 1. MNN models estimate the values of TIPs representing musical image for transformation of original theme music. In this paper, TIPs consist of some pairs of adjectives that are selected referring to a study that retrieves many genres musical works with pairs of adjectives representing musical image [10]. These are *happy-sad*, *heavy-light*, *hard-soft*, *stable-unstable*, *clear-muddy*, *calm-violent*, *smooth-rough*, *thick-thin*. Preexperiments are performed in order to confirm which pairs of adjectives are necessary for TIPs. The procedures of the preexperiments are as follows. (1) Fixing musical instruments, tempos, tonalities, tones, chords in a melody part and accompaniment parts patterns at random, 125 variations are generated. (2) Some subjects, who have no experience to play some musical instruments over 3 years, listen to the variations and express impressions on them with 8 pairs of adjectives. (3) If the subjects feel that it is difficult to evaluate the difference among the variations with some pairs of adjectives, they give the pairs. The results of the pre-experiments show that it is difficult to evaluate the difference among the variations using adjectives *hard-soft*, *stable-unstable*, *smooth-rough*, or *thick-thin*. Then, in this paper these four pairs of adjectives are not used. That is, four pairs of adjectives, which are parameters on *degree of change from original theme music* shown in Table 2, are used. Each parameter value is a real number in [0.0, 1.0].

Table 2: Transformation image parameters.

Parameters	Values
Calm-violent	[0.0–1.0]
Heavy-light	[0.0–1.0]
Happy-sad	[0.0–1.0]
Clear-muddy	[0.0–1.0]
Degree of change from original theme music	[0.0–1.0]

The MIA section estimates the values of TIPs from information on a picture scene. In generation of variations on theme music fitting to story scenes, information on story scenes necessary for the estimation of the values of TIPs is dependent on media representing a story, for example, pictures, texts or animations or the contents of a story, for example, a serious story, a story for children, and is not determined uniquely. Therefore, it is necessary to consider the selection of information on picture scenes for the estimation of the values of TIPs. However, since in this paper, input to the present system is limited to information on pictures scenes, the paper does not discuss this point. In the future it is necessary to change information according to media representing a story or the form of a story.

4. THEME MUSIC TRANSFORMATION (TMT) SECTION

4.1. Procedure on generation of variations [5]

Inputs to the TMT section are original theme music and values of TIPs obtained by the MIA section, and outputs are MIDI files of variations on theme music. MIDI files consist of the melody part and six accompaniment parts. The accompaniment parts consist of an obbligati part, a backing parts 1 and 2, a bass part, a pad part, and a drums part. The TMT section modifies impressions of inputted original theme music varying the following components of MIDI files [5]: (1) scores of melody parts, (2) tempos, (3) tonalities, (4) accompaniment patterns of accompaniment parts, and (5) tones.

4.2. Structure of TMT section

The TMT section transforms given original theme music according to inputted TIPs and outputted sets of MIDI-formatted candidates of variations on given theme music as shown in Figure 2. GAs are applied to the transformation of

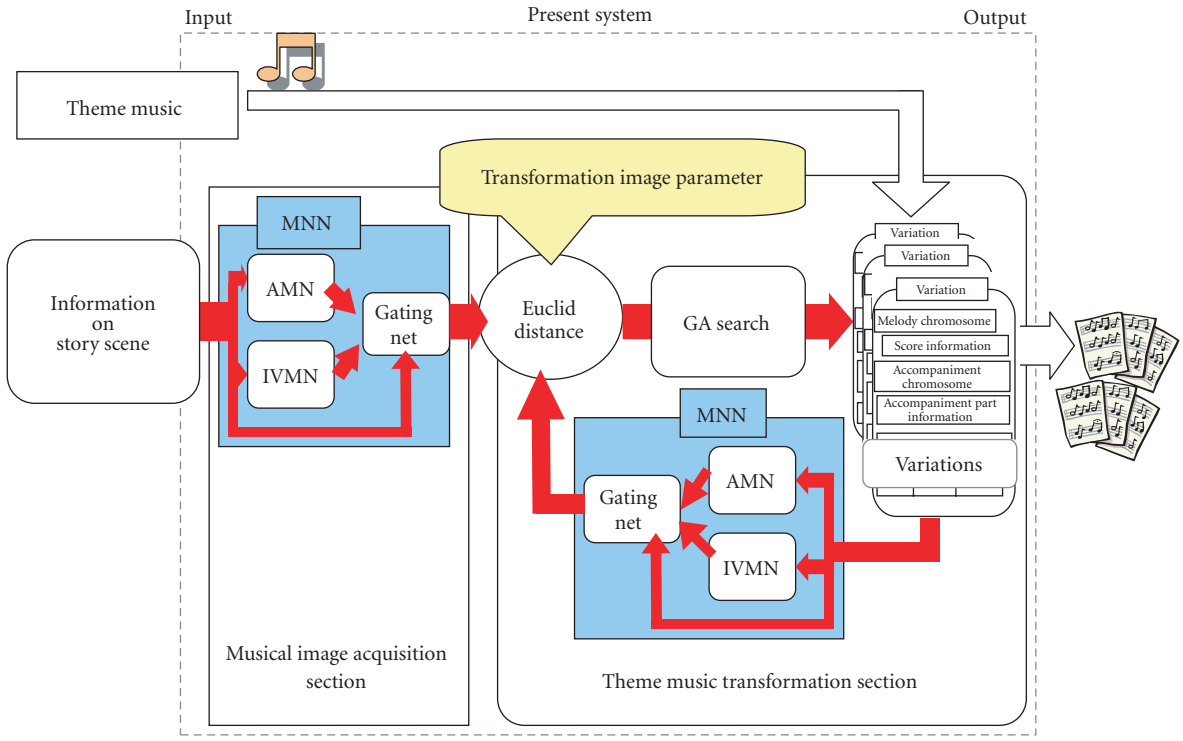


Figure 1: System structure.

a given theme music fitting to TIPs, where a variation generated from a given theme music is represented by a chromosome in the framework of GAs. In this paper, GAs parameters are abbreviated as follows.

(1) N: Population size
(2) T: Maximum number of generations
(3) N_{new}: The number of individuals generated randomly
(4) N_{user}: Partial population size presented to user
(5) P_c: Crossover probability
(6) P_m: Mutation probability.

Procedures in the TMT section are as follows.

(1) N variations are generated from inputted theme music in the form of chromosomes.
(2) Fitness values of chromosomes are calculated according to the inputted values of TIPs and melodies of original theme music.
(3) GAs operations of crossover and mutations are performed. Next generation population is generated. Go back to step (2).

4.2.1. Structure of chromosome

Variations consist of three kinds of chromosomes such as *Melody Chromosome*, *Accompaniment Chromosome*, and *Status Chromosome*.

The melody chromosome has melody part score information. Melody part score information is represented by the format shown in Figure 3. A given original theme music is represented as an initial chromosome. The accompaniment chromosome has accompaniment part information. The playing pattern number and the performance type of the obbligati part in the accompaniment part are represented by chromosomes, where each information is represented with 1 byte as shown in Figure 3. Initial chromosomes have random values for information. The status chromosome has information on a tempo, a tonality, and a tone. Tempo, tonality, melody part tone, and obbligato part tone are represented by a chromosome as shown in Figure 3. Tempo (60–200 [BPM]), tonality (a major scale or minor one), and tone are also represented with 1 byte. Initial chromosomes have random values for information.

4.2.2. Calculation of fitness value [5]

Fitness values of chromosomes are calculated according to the inputted values of TIPs and melodies of original theme music [5]. Let i ($i = 1, 2, \ldots, N$) be the chromosome number, that is, the variation number, and Fitness_i represents the fitness value of the ith variation. Fitness_i is defined as

$$\text{Fitness}_i = \text{Melody Fitness}_i + \text{Impression Fitness}_i, \quad (1)$$

where Melody Fitness_i is the fitness value of score information in the melody part of the ith variation referring to [11], and $\text{Impression Fitness}_i$ is the fitness value of impressions on the ith variation [5]. Impression values of variations are estimated by MNN models. These impression values are degrees of four pairs of adjectives used in TIPs estimation. MNN models are obtained by the relation between feature spaces of variations and impression values.

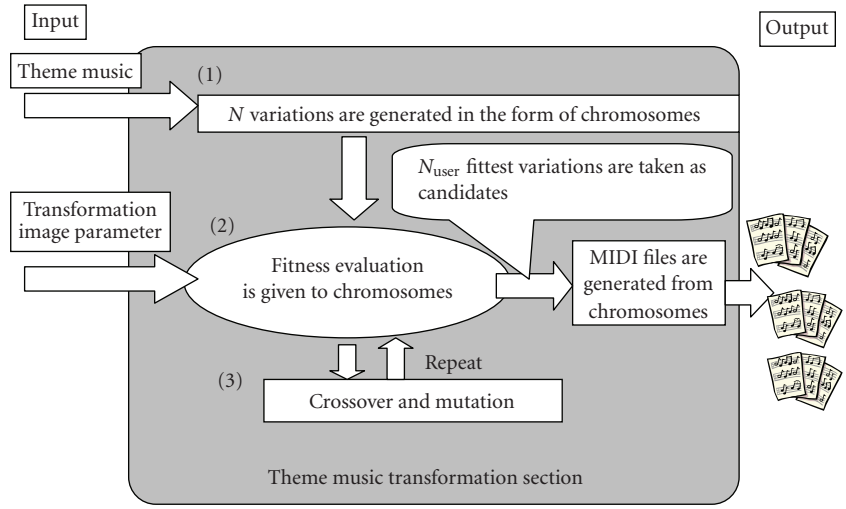


Figure 2: Theme music transformation section.

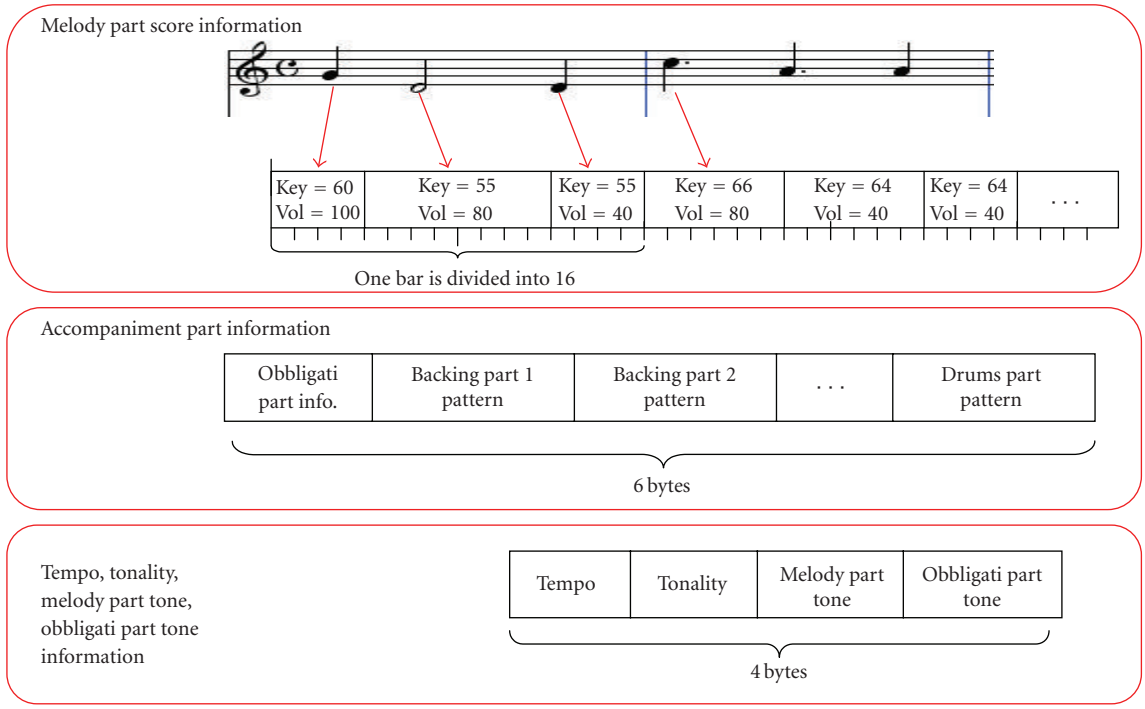


Figure 3: Three kinds of chromosomes.

Smaller the value of Fitness_i is, the better the ith variation is. Procedures of calculation of fitness values are shown in Figure 4.

4.2.3. GA operations

$(N - N_{\text{new}})$ individuals of parent candidates are selected by the tournament selection according to the fitness values obtained in 4.2.2. Crossovers at probability of P_c and mutations at P_m are applied to parent candidates. N_{new} individuals are generated at random. Crossover and mutation are performed as follows.

Crossover

uniform crossover is applied to melody chromosomes obtained by the generative theory of total music grouping structure analysis [12] in every group.

Mutation

random values are assigned to the accompaniment chromosome and the status chromosome. Varying score information on the melody part described in [5] is applied to melody chromosomes.

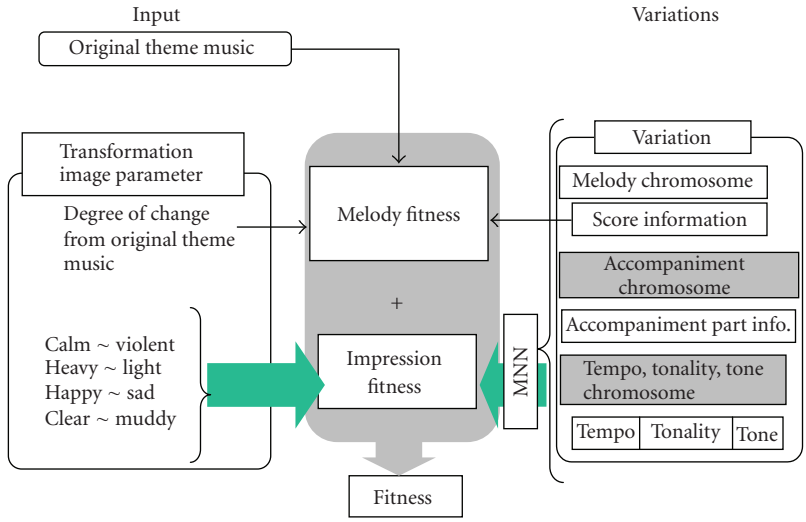


Figure 4: Calculation of fitness values.

5. MNN STRUCTURE

The present system uses MNN models to represent (1) relations between story scenes and values of TIPs in the MIA section, and (2) relations between features of variations and musical impressions in the TMT section. MNN models in the present system consist of AMN, IVMN, and the gating network as shown in Figure 5. When the present system adjusts its MNN models for each user, IVMN and the gating network are obtained by learning of user's data of individual variation of feeling of music and stories.

AMN is a hierarchical neural network model which consists of sigmoid neurons. AMN is constructed using questionnaire data of subject's feelings for music and stories. The questionnaire data are obtained referring to [6].

IVMN is a hierarchical neural network model which consists of RBF neurons. RBF is a function responding to input values in a local area. Therefore, an RBF network is easy to be adjusted online and fast. When a user is not satisfied with outputs of MNN, learning data of IVMN are generated and saved in the present system. Input values of learning data are input values of MNN. Output values of learning data are evaluation values by each user.

The gating network is an RBF network switching over between AMN and IVMN. The gating network judges whether input values of MNN are close to the area learned by IVMN or not. When a user is not satisfied with outputs of MNN, learning data of IVMN are generated and saved in the present system. Learning data of the gating network are input values of MNN. Output values of learning data of the IVMN are evaluation values by users.

IVMN and the gating network are constructed by the method proposed in [13] using all data saved in the present system.

Outputs of MNN models are defined as

$$f_{\mathrm{MNN}}(x) = \begin{cases} f_{\mathrm{personal}}(x) : g(x) \geq t \\ f_{\mathrm{average}}(x) : g(x) < t, \end{cases} \quad (2)$$

where $g(x)$ is an output value of the gating network $f_{\mathrm{personal}}(x)$ is an output value of the IVMN $f_{\mathrm{average}}(x)$ is an output value of the AMN, and t is a threshold of switching AMN and IVMN.

6. EXPERIMENTS

Experiments are performed to evaluate the present system by 8 undergraduate/graduate students. In the experiments, GA parameters are set at the following values: $N = 100$, $T = 100$, $N_{\mathrm{user}} = 3$, $N_{\mathrm{new}} = 20$, $P_c = 70\%$, $P_m = 20\%$. In the experiments, the threshold of switching AMN and IVMN by a gating network is set at 0.75. Musical works are chosen at random from prepared seventeen MIDI files of classical tunes or folk tunes, and are used as theme music of stories.

6.1. Construction of IVMN and gating network

IVMN and the gating network for each subject are constructed in the following procedures.

(1) Story scenes and theme music are inputted to the present system. The present system generates N variations according to each story scene and outputs them.
(2) When a subject is satisfied with one of outputted variations, go to (8). When a subject is not satisfied with any variations, go to (3).
(3) A subject looks at the values of TIPs estimated by the present system. The values of TIPs are presented to a subject in the form of Figure 6.
(4) The present system adjusts MNN models according to two cases as shown in Figure 7. That is, a subject feels (a) presented musical image is not suitable for story scenes or (b) generated variations are different from presented musical images.
    (a) When a subject feels that presented musical image is not suitable for story scenes, a subject evaluates whether the values of TIPs fit to story

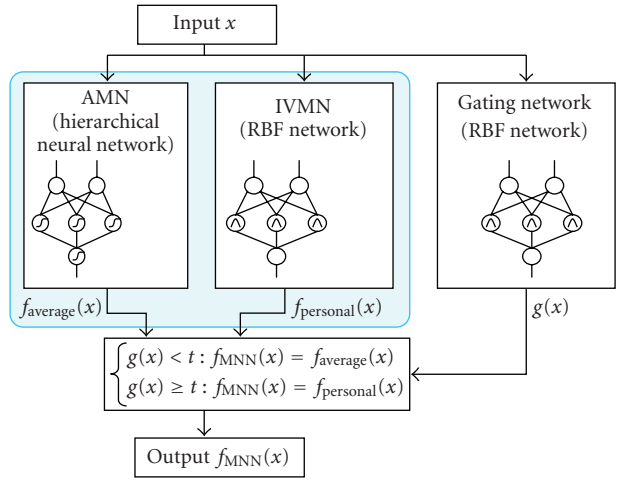


Figure 5: Concept figure on MNN.

scenes by the interface shown in Figure 6. Variations are generated according to values of TIPs evaluated by a subject. Go to (5).

(b) When a subject feels generated variations are different from presented images, a subject chooses one of$N$ variations. A subject evaluates his/her impressions using the 7-point scale method shown in Figure 8, where evaluation items are pairs of the same adjectives as the ones used as TIPs estimation. In this procedure the human interface shown in Figure 8 is used. Variations are generated by using modified MNN models. Go to (6).

(5) A subject listens to N variations. When a subject is satisfied with one of outputted variations, go to (7). When a subject is not satisfied with any variations, go to (3).

(6) A subject listens to N variations. When a subject is satisfied with one of outputted variations, go to (8). When a subject is not satisfied with any variations, go to (3).

(7) MNN models in MIA section are adjusted by the relation between information on story scenes and values of TIPs evaluated by a subject.

(8) Go to an evaluation of the next scene.

6.2. Experiment 1

Three story scenes are inputted into the present system and variations are generated, where the story scenes are different from the ones used in Section 6.1 and MNN models are adjusted for each subject in Section 6.1. This experiment confirms whether the present system generates variations on theme music reflecting subject's feeling of music and stories.

Let twelve story scenes be $S_i (i = 1, \ldots, 12)$. A subject is asked to read S_i and to evaluate musical images fitting S_i using the 7-point scale method (e.g., (7) very calm through (1) very violent), where evaluation items are 4 pairs of the same adjectives as the ones used in TIPs estimation. Let the evaluation values of S_i by a subject be $I_{S_i} = (a_{i1}, a_{i2}, a_{i3}, a_{i4})$ and let twelve variations generated from S_i by the present system be $P_i (i = 1, \ldots, 12)$. A subject is asked to evaluate impressions of P_i by 7-point scale method, where evaluation items are four pairs of the same adjectives as the ones used in TIPs estimation, and P_i are presented to a subject at random. Let impressions of P_i evaluated by a subject be $I_{P_i} = (b_{i1}, b_{i2}, b_{i3}, b_{i4})$, where a_{i1} and b_{i1}, a_{i2} and b_{i2}, a_{i3} and b_{i3}, and a_{i4} and b_{i4} are evaluation values of "violent-calm," "heavy-light," "clear-muddy," and "sad-happy," respectively. These variables have integer values in [1, 7] evaluated by a subject. In this experiment, cosine correlations [14] between I_{S_i} and I_{P_i} are used for the evaluation whether the generated variations are reflecting subject's feelings for music and stories or not. Cosine correlation $\text{Sim}(I_{S_i}, I_{P_i})$ is defined as

$$\begin{aligned} \text{Sim}(I_{S_i}, I_{P_i}) &= \cos\left(\arg\left(I_{S_i}, I_{P_i}\right)\right) \\ &= \frac{\sum (a_{ij} \times b_{ij})}{\sqrt{\sum (a_{ij})^2} \times \sqrt{(b_{ij})^2}}, \quad (1 \leq j \leq 4), \end{aligned} \tag{3}$$

when $\text{Sim}(I_{S_i}, I_{P_i})$ is close to 1.0, generated variations are reflecting users feelings for music and story well.

6.3. Result 1

$\text{Sim}(I_{S_i}, I_{P_i})$ are shown in Table 3. It is found that 80% of the whole of $\text{Sim}(I_{S_i}, I_{P_i})$ is 0.9 or more, and the present system is able to generate variations reflecting subject's feelings to music and stories.

6.4. Experiment 2

Other three story scenes are inputted into the present system and variations are generated, where MNN models in the present system are adjusted for each subject in Section 6.1. A subject is asked to evaluate with 7-point scale method whether variations on theme music fit impressions of each presented story scene or not; (7) very suitable (6) suitable (5)

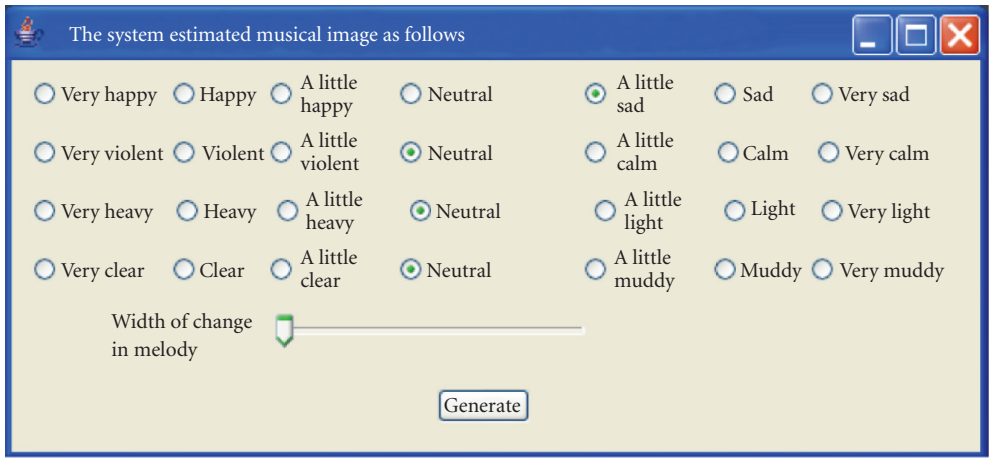


Figure 6: TIPs estimation by present system.

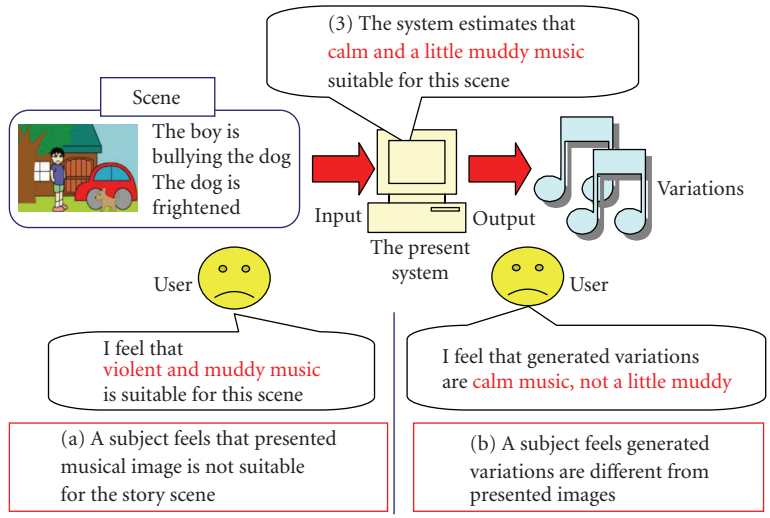


Figure 7: MNN models adjustment.

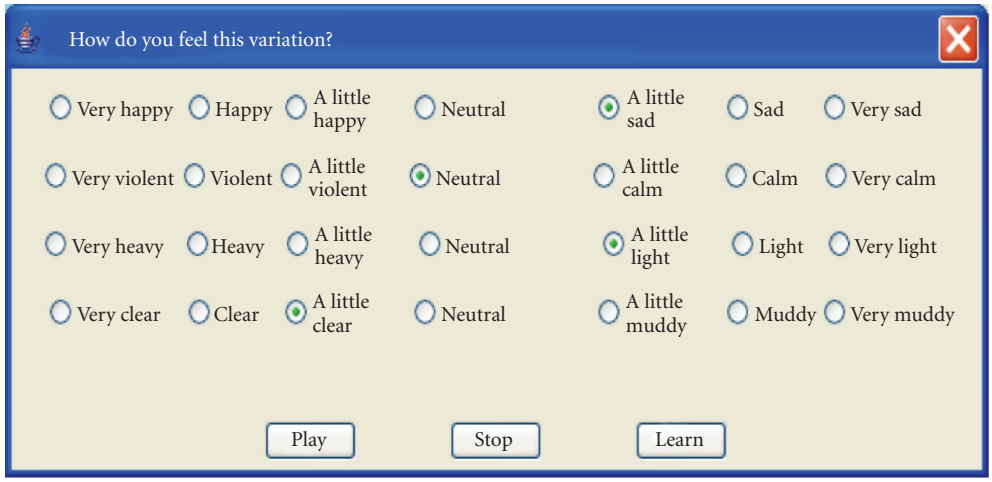
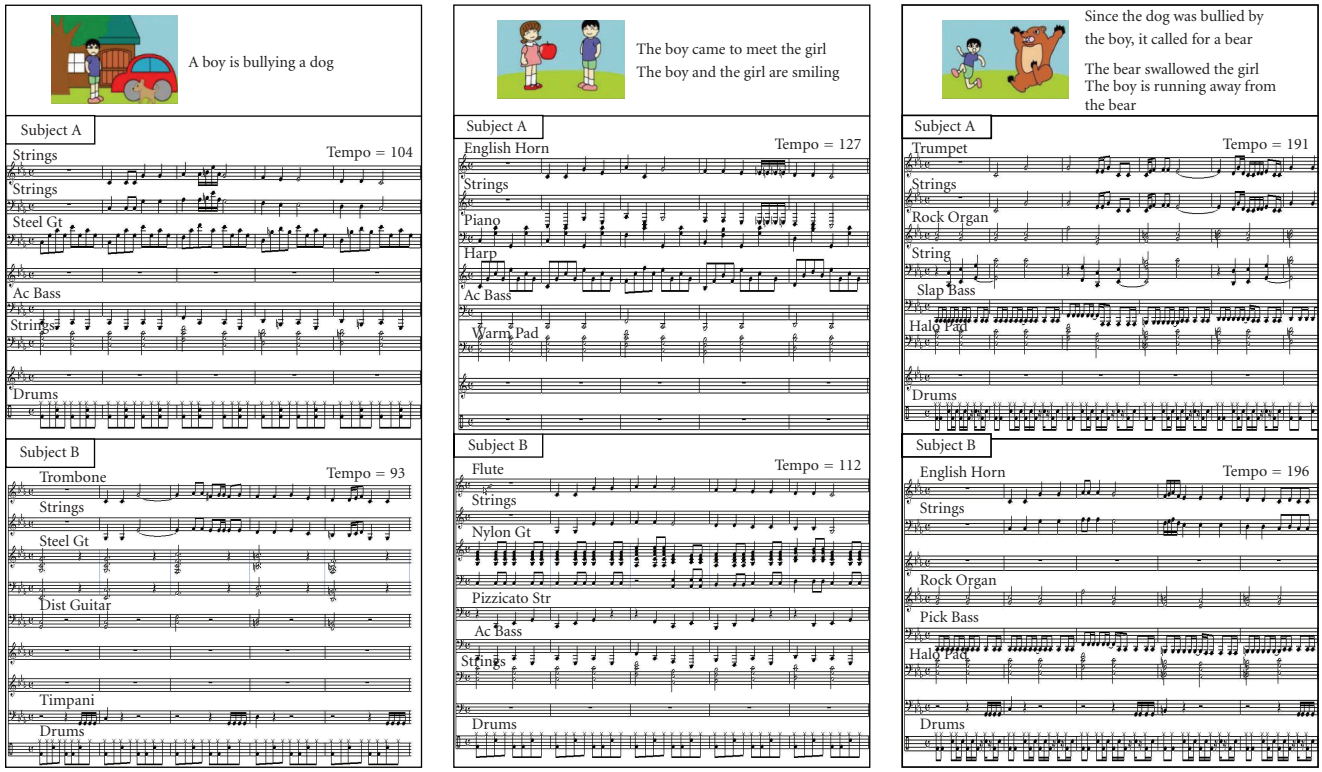


Figure 8: Interface.

Graph	Average of evaluation values	Decentralization of evaluation values
Average system	4.97	1.82
Present system	5.69	1.06

Figure 9: Distributions of evaluation values.



(a) (b) (c)

Figure 10: Examples of variations.

a little suitable (4) neutral (3) little suitable (2) not suitable (1) not suitable at all. Furthermore, to confirm the effectiveness of IVMN, variations generated by the present system and the ones generated by an average system are compared with each other, where the average system is the system whose IVMN in MNN models are average among subjects.

6.5. Result 2

Experimental results are shown in Table 4. It is found that the present system gets evaluation (5) or (6) or (7) in approximately 87.5% of all evaluation results.

Distributions of evaluation values of variations generated by the present system and those of variations generated by the average system are shown in Figure 9. It is found that variations generated by the present system are evaluated higher in the large percentage than variations generated by the average system. An average and a decentralization of evaluation values are also shown in Figure 9. The decentralization of evaluation values of variations generated by the present system is lower than that of variations generated by the average system. It is found that constructing IVMN and the gating network for each user are effective.

Figure 11: Theme music.

Figure 10 shows examples of the variations on theme music by subjects A and B, where original theme music is Twinkle Stars as shown in Figure 11. From these figures it is found

Table 3: Cosine Similarity.

Story	Scene	Subject							
		G	H	I	I	K	L	M	N
1	1	0.86	0.94	0.93	0.95	0.96	0.98	0.99	0.97
	2	0.83	0.99	0.98	0.97	0.98	0.99	0.98	0.97
	3	0.94	0.99	0.82	0.95	0.95	0.99	0.98	0.94
	4	0.96	0.90	0.96	0.96	0.96	0.95	0.98	0.94
2	1	0.84	0.97	0.97	0.94	0.94	0.95	0.94	0.89
	2	0.91	0.97	0.99	0.85	0.99	0.99	0.99	0.98
	3	0.98	0.95	0.98	0.93	0.93	0.97	0.98	0.95
	4	0.97	0.92	0.96	0.82	0.91	0.97	1.00	1.00
3	1	0.84	0.94	0.83	0.95	0.85	0.98	0.98	0.97
	2	1.00	0.98	0.99	0.96	0.98	0.99	0.99	0.99
	3	0.99	0.93	0.98	0.94	0.95	0.93	0.88	0.40
	4	0.65	0.98	0.79	0.80	0.98	0.97	0.94	0.50
	Average	0.90	0.96	0.93	0.92	0.95	0.97	0.97	0.88

Table 4: Evaluation Values.

Story	Scene	Evaluation Values							
		G	H	I	I	K	L	M	N
1	1	6	6	5	4	5	6	3	5
	2	6	7	7	6	5	5	5	4
	3	5	7	7	6	6	7	6	6
	4	6	6	6	7	3	7	6	4
2	1	7	7	4	6	5	7	4	6
	2	6	5	6	5	6	6	6	5
	3	6	6	7	6	6	7	6	5
	4	7	6	7	7	4	7	6	6
3	1	7	7	6	5	5	5	4	5
	2	7	6	5	6	6	6	4	5
	3	5	6	6	7	4	5	6	5
	4	7	6	6	6	4	6	5	3

that although the same scenes are given to the subjects, various theme tunes are transformed by the present system.

Variations on theme music generated by the present system are dependent on subjects' impressions on story expressed by pictures. Therefore, even if the same pictures are given, generated variations are different among subjects. Nevertheless, subjects themselves are satisfied with generated variations. Then it is found that the present system generates variations on theme music fitting to subjects' impressions on story well. However, subjects' impressions on story usually change according to time and environment in which subjects are. The present system does not deal with the variations depending on these factors, time, environment, and so forth. This is a future work.

7. CONCLUSIONS

This paper presents the system which transforms a theme music fitting to story scenes represented by texts and/or pictures, and generates variations on the theme music. The present system varies (1) melodies, (2) tempos, (3) tones, (4) tonalities, and (5) accompaniments of a given theme music based on impressions of story scenes using neural network models and GAs. Differences of human's feeling of music/stories are important in multimedia content creation. This paper proposes the method that adjusts the models in the present system for each user. The results of the experiments show that the system transforms a theme music reflecting user's impressions of story scenes.

REFERENCES

[1] Z. Iwamiya, *Multimodal Communication on Music and Visualizations*, Kyushu University Press, Fukuoka, Japan, 2000.

[2] S. Takahasi, M. Okamoto, and H. Ohara, "Voice and sound processing technology for easy, comfortable, convenient communications environment," *NTT Technical Journal*, pp. 8–9, 2004 (Japanese).

[3] H. Liu, H. Lieberman, and T. Selker, "A model of textual affect sensing using real-world knowledge," in *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*, pp. 125–132, Miami, Fla, USA, January 2003.

[4] H. Takagi and T. Noda, "Media converter with impression preservation using a neuro-genetic approach," *International Journal of Hybrid Intelligent Systems*, vol. 1, no. 1, pp. 49–56, 2004.

[5] K. Ishizuka and T. Onisawa, "Generation of variations on theme music based on impressions of story scenes," in *Proceedings of the International Conference on Games Research and Development*, pp. 129–136, Perth, Western Australia, December 2006.

[6] Y. Kiyoki, T. Kitagawa, and T. Hayama, "A metadatabase system for semantic image search by a mathematical model of meaning," *ACM SIGMOD Record*, vol. 23, no. 4, pp. 34–41, 1994.

[7] W. Apel, *Harvard Dictionary of Music*, Harvard University Press, London, UK, 2nd edition, 1973.

[8] S. Kato and T. Onisawa, "Generation of consistent linguistic expressions of pictures," *Journal of Japan Society for Fuzzy Theory and Intelligent Infomatics*, vol. 17, no. 2, pp. 233–242, 2005 (Japanese).

[9] K. Watanabe, *Neural Network Computational Intelligence*, Morikita Press, Tokyo, Japan, 2006.

[10] T. Ikezoe, Y. Kazikawa, and Y. Nomura, "Music database retrieval system with sensitivity words using music sensitivity space," *Journal of Japan Information Processing Society*, vol. 42, no. 12, pp. 3201–3212, 2001 (Japanese).

[11] T. Kadota, M. Hirao, A. Ishino, M. Takeda, A. Shinohara, and F. Matsuo, "Musical sequence comparison for melodic and rhythmic similarities," in *Proceedings of the 8th International Symposium on String Processing and Information Retrieval (SPIRE '012001)*, pp. 111–122, Laguna De San Rafael, Chile, November 2001.

[12] F. Lerdahl and R. Jackendoff, *A Generative Theory pf Tonal Music*, MIT Press, Cambridge, Mass, USA, 1983.

[13] A. Hattori and H. Nakayama, "Additional learning and active forgetting by support vector machine and RBF networks," Tech. Rep., Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, 2002.

[14] A. Yamaue and S. Kurachi, *Psychological Statistics*, Kitaoozi Press, Tokyo, Japan, 1991.