A Natural Language Dialogue System for Impression-based Music Retrieval

Tadahiko Kumamoto

Abstract—Impression-based music retrieval is the best way to find pieces of music that suit the preferences, senses, or mental states of users. A natural language interface (NLI) is more useful and effective than a graphical user interface for impression-based music retrieval since an NLI interprets users' spontaneous input sentences to represent musical impressions and generates query vectors for music retrieval. Existing impression-based music retrieval systems, however, have no dialogue capabilities for modifying the most recently used query vector. We evaluated a natural language dialogue system we developed that deals not only with 164 impression words but also with 14 comparative expressions, such as "a little more" and "more and more," and, if necessary, modifies the most recently used query vector through a dialogue. We also evaluated performance using 35 participants to determine the effectiveness of our dialogue system.

Index Terms—Music retrieval, impression-based, natural langauge dialogue.

I. INTRODUCTION

WHEN users want to locate specific music data from the huge volumes contained in music databases, they usually input bibliographic keywords, such as title and artist. When they do not have any bibliographic keywords, they can use content-based music-retrieval systems that enable them to find data by singing the song, typing parts of the lyrics, or humming the tune [1], [2], [3], [4]. However, these systems are ineffective if they do not specify the exact music data they want to find. In such situations, impression-based music-retrieval systems are best because they enable users to find pieces of music that suit their preferences, senses, or mental states [5], [6], [7], [8], [9].

Information for impression-based music-retrieval systems is generally entered using one of the following three methods: (i) users select one or more impression words from the multiple words presented [5], [8], (ii) users select one or more impression words from the multiple words presented and evaluate each of the selected words using a Likert scale [6], and (iii) users select one or more pairs of impression words from the multiple pairs presented and evaluate each of the selected pairs using a Likert scale [7], [9]. With these approaches, increasing the numbers of words presented increases the cost to the user in terms of time and the labor required to input impressions. A set of words that is

too limited, on the other hand, will often not allow users to accurately represent their target impressions. A natural language interface (NLI), therefore, is needed so that users can input impressions without consciously limiting their vocabulary and wording.

Few NLIs have so far been developed as a user interface for an impression-based multimedia content retrieval system. For example, we have developed an NLI that interprets 164 impression words, such as "happy" and "sad," and 119 degree modifiers, such as "a little" and "comparatively," and generates a query vector for music retrieval [10], [11]. However, our NLI does not have dialogue capabilities for modifying the most recently used query vector through a dialogue. That is, when a user's impression from a listening to a retrieval result is not similar to or does not match the user's inputted impressions, the user is asked to enter a new sentence to obtain more accurate retrieval results rather than modify the most recent query vector. Harada et al. have developed an impression-based image-retrieval system with an NLI that deals with 40 impression words and a few comparative expressions [12]. However, their NLI only understands a few stereotypical expressions, such as "I would like to have a simpler one," and does not have sufficient dialogue capabilities for interactive impression-based retrieval.

In this paper, we, therefore, describe a natural language dialogue system we developed that helps users to generate or modify query vectors for impression-based music-retrieval in Japanese. We modified our impression-based music-retrieval system to include an NLI [9], [10], [11] as a base module for our dialogue system since our NLI deals with many impression words but has no dialogue capabilities for interactive retrieval. That is, we developed a method for dealing with 14 comparative expressions, and then incorporated it into our NLI. This enables users to obtain more accurate retrieval results for impression-based music retrieval since users can easily modify the most recently used query vectors through a dialogue with our dialogue system. Note that such comparatives as "brighter" and "happier" are not used in the target language of Japanese, while adjectives modified by such comparative expressions as "more," "a little more," and "still more" are used.

The remainder of this paper is organized as follows. Sect. II describes the specifications of the query vectors that are valid in our impression-based music-retrieval system. Sect. III proposes a method for interpreting users' spontaneous input sentences and, if necessary, modifying the most recently

TABLE I
TEN IMPRESSION SCALES FOR REPRESENTING MUSICAL IMPRESSIONS.

N	Impression word pair	N	Impression word pair
1	Quiet — Noisy	6	Leisurely — Restricted
2	Calm — Agitated	7	Pretty — Unattractive
3	Refreshing — Depressing	8	Happy — Sad
4	Bright — Dark	9	Relax — Arouse
5	Solemn — Flippant	10	The mind is restored — The mind is vulnerable

TABLE II
PORTION OF INTERPRETATION RULES BETWEEN IMPRESSION WORDS AND QUERY VECTORS.

Scale No.	1	2	3	4	5	6	7	8	9	10
pitiful	nil	nil	2.48	2.13	nil	nil	nil	1.75	nil	nil
not pitiful	nil	nil	nil	5.41	nil	nil	nil	5.37	nil	nil
classic	5.42	5.56	nil	3.47	5.57	nil	5.51	nil	5.06	5.09
gentle	5.49	5.79	5.62	5.27	nil	5.62	6.01	5.10	5.85	6.16
powerful	2.13	nil	nil	nil	nil	nil	nil	nil	2.38	nil

TABLE III
PARAMETERS USED TO INTERPRET 14 COMPARATIVE EXPRESSIONS. SOME COMPARATIVE EXPRESSIONS WERE TRANSLATED INTO THE SAME ENGLISH EXPRESSIONS. HOWEVER, THERE IS A SMALL DIFFERENCE BETWEEN THEIR NUANCES IN JAPANESE. A STANDS FOR THE ADJUSTED COEFFICIENT OF DETERMINATION.

Comparative expression	Coefficient b	Constant c	A
a little more (mousukoshi)	0.92	0.92	0.998
more; longer; farther; —er (motto)	0.89	1.36	0.998
more; longer; farther; —er (yori)	0.91	1.15	0.996
a little more (mosukoshi)	0.95	0.60	0.995
more and more; increasingly (masumasu)	0.88	1.53	0.994
a little more (mosotto)	0.92	0.87	0.994
a little bit (honno-sukoshi)	0.96	0.44	0.989
much more; all the more (issou)	0.78	2.22	0.959
still more; further (sarani)	0.88	1.54	0.953
still more; further (ichidanto)	0.80	2.28	0.948
far; much (zutto)	0.85	2.26	0.916
far; much (zuutto)	0.40	4.86	0.906
with a jerk; much better (gutto)	0.48	4.26	0.619
with a jerk; much better (gunto)	0.39	4.93	0.535

used query vectors. Sect. IV shows the process flow of our dialogue system into which the proposed method was incorporated. Sect. V contains the details and the results of a performance-evaluation experiment with 35 participants. Finally, Sect. VI concludes this paper and describes the future works to be undertaken.

II. REQUIRED QUERY VECTOR SPECIFICATIONS

This section contains a description of the specifications for query vectors that are valid in the impression-based music-retrieval system into which our proposed method is incorporated.

A query vector has ten components. Each component corresponds sequentially to each of the ten impression scales listed in Table I. A component's value is a real number between 0 and 8 corresponding to a seven-step Likert scale, and symbol "nil" denotes a "don't care" term in an impression scale defined using paired impression words. For instance, impression scale No. 8 "Happy — Sad" is characterized by the seven categories including "very happy," "happy," "a little happy," "medium," "a little sad," "sad," and "very sad," that correspond to 7.0, 6.0, 5.0, 4.0, 3.0, 2.0, and 1.0, respectively. That is, to find musical pieces that will create a happy impression, the following 10-dimensional vector would be generated as a query vector.

$$(nil\ nil\ nil\ nil\ nil\ nil\ nil\ 6.0\ nil\ nil)$$

Similarly, to find dark and sad musical pieces, the following query vector would be generated.

$$(nil\ nil\ nil\ 2.0\ nil\ nil\ nil\ 2.0\ nil\ nil)$$

The NLI into which our proposed method of enabling interactive impression-based retrieval is incorporated generates query vectors from 164 impression words using the interpretation rules between impression words and query vectors as listed in Table II. For example, when impression word "gentle" is only extracted from the sentence a user entered, the following query vector will be generated by applying an interpretation rule to the impression word.

$$(5.49\ 5.79\ 5.62\ 5.27\ nil\ 5.62\ 6.01\ 5.10\ 5.85\ 6.16)$$

III. DIALOGUE CAPABILITIES FOR IMPRESSION-BASED RETRIEVAL

This section proposes a method to interpret sentences for modifying the most recently used query vector, which enables interactive impression-based retrieval.

First, we defined comparative expressions that our dialogue system should interpret for interactive impression-based retrieval. We extracted 14 adverbs and adverbial expressions representing the complete degree of change from a Japanese thesaurus called "Ruigo-Kokugo-Jiten" [13] and set them as target comparative expressions in this paper. All target comparative expressions are listed in Table III together with the parameters used to interpret the comparative expressions, where the parameters will appear in the equation (2).

Next, we conducted a questionnaire-based experiment using 50 women and 50 men to investigate how a query vector, $v_i(i = 1, 2, \cdots, 10)$, should be modified on the basis of comparative expressions extracted from a sentence entered to modify the query vector. In this experiment, we assumed that a specified query vector had already been used. The query vectors to be generated in finding "totemo-shizukana (very quiet)," "shizukana (quiet)," "sukoshi-shizukana (a little quiet)," "dochiratomo-ienai (medium)," "sukoshi-shizukadenai (not a little quiet)," "shizukadenai (not quiet)," or "totemo-shizukadenai (not quiet at all)" tunes were assumed and used. Under each
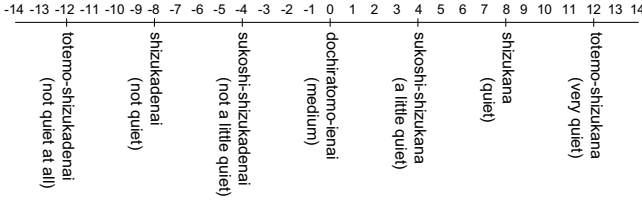
Fig. 1. A base scale for scoring. Participants marked their evaluation on the base scale.
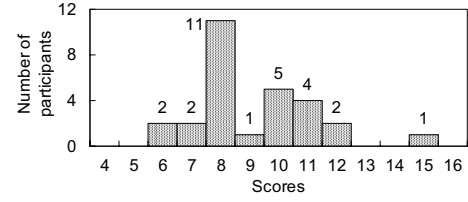

Fig. 2. How much "ichidanto-shizukana (still quieter)" tunes should be to create a quiet impression were marked on the base scale by twenty eight participants under the assumption that "sukoshi-shizukana (a little quiet)" tunes were the most recently retrieved. This histogram shows the results.

assumption, about 30 participants evaluated impressions of the phrase containing an impression word and a comparative expression using a base scale for scoring, as shown in Fig. 1. For example, if a query vector for finding a "shizukana (quiet)" tune was just used, the participants were then asked to mark how much "motto-shizukana (quieter)" tunes should be to create a quiet impression on the base scale printed on a sheet of paper.

We found that the estimation results in this experiment varied widely, as shown in Fig. 2. We, therefore, introduced a method to determine the representative value of a histogram. The process flow is outlined in Fig. 3. First, if the number of participants in the mode of an input histogram is more than the majority, the method determines the value of the mode as a representative value of the histogram and terminates the process. Otherwise, the method performs the following processes. (1) Simple moving averages (*MA*) are first computed with the range of five for an input histogram. Then, *MA* is computed using the following equation in this paper.

$$MA_{score} = \frac{1}{5} \sum_{i=-2}^{2} N_{score+i} \qquad (1)$$

where N_x denotes the number of participants in the score, x. (2) The number, N, of participants in the range between a score with the maximum moving average plus minus d is computed when the initial value of d is 1. (3) If N is greater than the majority, the mean value computed only from the scores of the participants in the range is determined as a representative value of the histogram, and the process is terminated. Otherwise, d is increased by 1, and the process then proceeds to Step (2).

Next, we used regression analysis to obtain a regression equation or linear equation between the most recently used query vector, w_i, and a query vector modified by a sentence that included a comparative expression, w_i'. Note that w_i equals the value of the ith component of the query vector assumed to have been used most recently, and w_i' equals the representative value subsequently obtained from the corresponding histogram using the method mentioned above. Since this equation was computed based on the 29-step base scale, the equation was converted in terms of the seven-step impression scale. Coefficient b and constant c of equation $(v_i' = bv_i + c)$ that were subsequently obtained were then
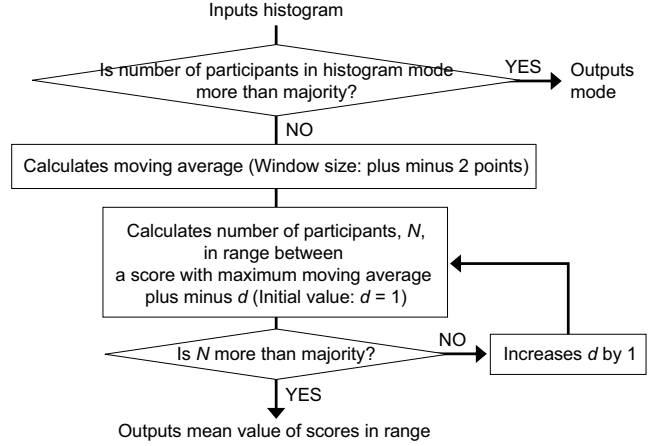

Fig. 3. Process flow for determining a representative value from the input histogram.

registered in a parameter table for interpreting comparative expressions, which is listed in Table III along with the corresponding adjusted coefficients of determination [14]. Where v_i and v_i' denote comparative expressions converted from w_i and w_i', respectively. We find that all the comparative expressions excluding "gutto (much better)" and "gunto (much better)" have very high adjusted coefficients of determination, and the adjusted coefficients of determination for "gutto (much better)" and "gunto (much better)" are not so low. Hence, we can say that satisfactory results were obtained in this regression analysis.

We used "shizukana (quiet)" as the impression word modified by a comparative expression. Since this impression word corresponds to 6 points on the seven-step scale, the equation $(v_i' = bv_i + c)$ cannot be applied to impression words corresponding to 4 or lower points, such as "sad" and "dark." To solve this problem, we assumed that the difference, $v_i' - v_i$, between the most recently used query vector and a newly generated query vector was in proportion to query vector x_i generated from the impression word modified by a comparative expression. We formulated the following simultaneous equation.
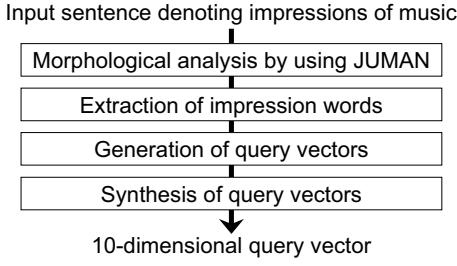
Input sentence denoting impressions of music

| Morphological analysis by using JUMAN |
| Extraction of impression words |
| Generation of query vectors |
| Synthesis of query vectors |

10-dimensional query vector

Fig. 4. Steps for generating query vectors from input sentences.

When $x_i = 6$, $v_i' = bv_i + c$
When $x_i = 4$, $v_i' = v_i$
When $x_i = 2$, $v_i' = v_i - (v_i' - v_i)$ $= (2-b)v_i - c$

By solving this simultaneous equation, we obtained the following equation.

$$v_i' = \frac{(bx_i - x_i - 4b + 6)v_i + c(x_i - 4)}{2}, \qquad (2)$$

where x_i denotes the value of the ith component of a query vector generated from the corresponding impression word, and the values for b and c are obtained from the parameter table listed in Table III. In sessions where $x_i = nil$, $v_i' = v_i$ and the value of v_i' is kept. In sessions where $x_i \neq nil$ and $v_i = nil$, the value of v_i is replaced with the value of the ith component from an impression vector [1] representing the first candidate of the most recently retrieved musical pieces.

IV. DIALOGUE CAPABILITIES FOR INTERACTIVE IMPRESSION-BASED RETRIEVAL

Sentences that users input into the NLI are processed as outlined in the steps shown in Fig. 4, and subsequent query vectors will be generated and used in inputs to retrieve music. Each step is outlined in the following.

A. Morphological Analysis

An input sentence is decomposed into words by using JUMAN [15], which is one of the most famous Japanese morphological analysis systems. The basic form, the part-of-speech name, and the conjugation name for each word are simultaneously annotated for the word as tags.

B. Extraction of Impression Words

Information for query vector generation, such as impression words, comparative expressions, and negative words, is extracted from annotated words obtained in the preceding step. If negative words were extracted, they would become a pair with the depending impression words for each pair, and each pair would be dealt with as one impression word in the following steps. For instance, the impression word "not-pretty" is extracted from the sentence "I don't want a pretty one."

[1] A 10-dimensional vector consisting of real values between 0.0 and 8.0 was automatically assigned to every musical piece using a function of the original impression-based music-retrieval system [9] we used in this paper.

C. Generation of Query Vectors

First, when the interpretation rules are applied to the impression words extracted in the preceding step, a query vector is generated from each of the words, as described in Sect. II. When there are results from most recently performed retrieval and a comparative expression is extracted with an impression word from an input sentence, the value of v_i' is obtained by substituting the values of v_i, b, c, and x_i into the equation (2), where v_i' denotes the value of the ith component of the query vector input to the music-retrieval system, v_i denotes the value of the ith component of the most recently used query vector, the values of b and c for the comparative expression are obtained from the parameter table listed in Table III, and x_i denotes the value of the ith component of the query vector generated from the impression word.

D. Synthesis of Query Vectors

If two or more query vectors are generated, the query vectors are synthesized by computing the mean value in each component, where "nil" is excluded from the calculation. For example, when the impression words "happy" and "gentle" are extracted from the input sentence, the following two query vectors are generated from the words.

(nil nil nil nil nil nil nil 6 nil nil)
(5.49 5.79 5.62 5.27 nil 5.62 6.01 5.10 5.85 6.16)

Then, these query vectors are synthesized, and the following query vector is obtained.

(5.49 5.79 5.62 5.27 nil 5.62 6.01 5.55 5.85 6.16)

V. PERFORMANCE-EVALUATION OF DIALOGUE CAPABILITIES

Nineteen women and sixteen men participated in a performance-evaluation experiment of the proposed dialogue capabilities. We first asked all 35 participants the question, "What impression should the musical pieces you want to obtain create?". Each of the participants replied to this question with four answers denoting her/his impressions of the musical pieces. We then asked the participants to individually enter their four answers into the impression-based music-retrieval system, which was proposed in the previous sections, and retrieve music [2] using the following procedures. When musical pieces were presented as retrieval results, the participants listened to the first retrieved musical piece with the shortest distance, and then evaluated whether the impressions

[2] The music database had 160 short classical pieces. The playback time ranged from about 30 seconds to several minutes. All the musical pieces have been released to the public through http://nocturne.vis.ne.jp/ and http://k2works.com/nerve/ and have been approved for secondary purposes. A 10-dimensional vector, consisting of real values between 0.0 and 8.0, was automatically assigned to each musical piece using a function of the original music-retrieval system [9] we used in this paper. The original music-retrieval system calculated the distances between a query vector and the vectors of musical pieces, and then presented, at most, the three musical pieces with the shortest distances.

Fig. 5. A captured snapshot of the user interface for retrieval. The text box in the upper part of the screen displays dialogue history, the text box in the left-middle part displays reports and instructions from the system, the text box in the right-middle part is for the user input, and the text box in the lower displays the results of the most recent retrieval. Rank, names of musical pieces, data size, distance, and so on are also displayed in this text box.

from this piece were similar to the input impressions using a five-point scale. For example, if the impressions from the first retrieved musical piece were very similar to those input, the participants would award five points. Conversely, if the impressions from the first retrieved musical piece were not at all similar to those input, the participants would award one point. For scores other than five points, the participants were asked to enter a sentence to modify the most recently used query vector and retrieve music again. The 14 comparative expressions, which the participants could enter, were printed on a sheet and presented to the participants, and the participants were asked to use the comparative expressions to form a sentence to modify a query vector. The participants repeated retrieving music until they awarded five points to the first retrieved musical piece or the results of the retrieval became empty. A sample of dialogues provided by our system is shown in Fig. 5, and the results of the experiment mentioned above are listed in Table IV.

Although the participants conducted a total of 140 music retrievals, they awarded five points in 51 (36.4%) of the 140 sessions. A score of five points was awarded for the first retrieval using a new query vector in 36 of the 51 sessions, and the scores were increased to five points by modifying the most recently used query vectors in 15 of the 51 sessions. Note that the remaining session of "Increased" was the case in which, although the score was increased to three points, the session was terminated for some reason.

Next, we analyzed the sessions in which a score of five points was not awarded at all. In Table IV, the "four points" for the new query vectors resulted from the fact that the

TABLE IV
PERFORMANCE-EVALUATION EXPERIMENT RESULTS.

Query vector	Score or change in score	Number of sessions
New	5 points	36
	4 points	1
	Number of zero hits	2
	Failed (out of target)	31
	Failed (use of unregistered word)	17
Modified	Increased	16
	No change	8
	Decreased	2
	Number of zero hits	13
	Failed (out of target)	9
	Failed (use of unregistered word)	5
Total		140

sentence to modify the most recently used query vector was interpreted as one for generating a new query vector due to a participant's mistype and the session was terminated. The reason why the situation in which the numbers of hits for new query vectors were zero occurred is because the negative form of suffix "sugiru (too)" [3] was not interpreted adequately. In total, 40 (28.6%) of the 140 sessions were classified "Failed (out of target)." This phenomenon occurred when the impression words, which the dialogue system could deal with, were not extracted from the sentences the participants entered. The impression words the dialogue system can deal with are limited to the 164 words that represent the affective characteristics of musical pieces or the change in the listeners' affective states. The input sentence the dialogue system failed to interpret contained critical comments about musical pieces (11 sessions), statements about scenic images (seven sessions), comments about a register of musical instrument (five sessions), and comments about musical structures (four sessions). We did not present the impression words the dialogue system could deal with to the participants to enable them to spontaneously enter sentences. In total, 22 (15.7%) of the 140 sessions were classified "Failed (use of unregistered word)." That is, any query vectors were not generated due to the use of words unregistered in the interpretation rules and sessions were terminated. This indicates the need to develop a method to deal with unregistered words and to incorporate it into the dialogue system. This is one of the areas for our future works. In 2 of the 140 sessions, the scores were decreased through dialogues. One case was when a participant scored low even though the same musical piece was presented as the first retrieval result, and the other was a case when a wrong retrieval result was presented due to wrong interpretation of the sentence indicating changed impressions during the performance of a musical piece. In 8 of the 140 sessions, the scores were not changed through dialogues and were less than five points. These cases occurred since the sessions were terminated for some reasons even though the conditions for

[3]The Japanese word "sugiru" is a suffix and forms a verb by being concatenated with an adjective and can be translated into the English word "too." An example usage is "it is too simple to do this."

finishing a session had not been satisfied. We would like to reconsider the instructions provided to participants.

As mentioned above, when the dialogue system successfully interpreted the user input sentences, the scores for musical pieces presented as the first retrieval results improved. This proves that the dialogue system enables users to obtain better retrieval results through dialogues for impression-based retrieval.

VI. CONCLUSION

We evaluated a method we proposed to deal with 14 comparative expressions and to enable interactive impression-based retrieval. We incorporated the proposed method into an existing impression-based music-retrieval system that lacked dialogue capabilities but accepted natural language input and conducted a performance-evaluation experiment with 35 participants. Results revealed that the dialogue system was effective in obtaining better retrieval results using dialogues than language input. Note that, although we did not mention in this paper the details of the 119 degree modifiers, such as "a little" and "comparatively" that an original natural language interface can deal with [11] due to page limits, our dialogue system also can deal with the 119 degree modifiers.

Performance evaluation of our dialogue system is the basis for our future work. Other future work includes developing a method to deal with unregistered words and a method to manage changed impressions of a musical piece during performance.

REFERENCES

[1] Ghias, A., Logan, J., Chamberlin, D., and Smith, B.: Query By Humming – Musical Information Retrieval in an Audio Database. In Proc. of ACM Int. Multimedia Conf., San Francisco, USA (1995)

[2] Blackburn, S. G., and De Roure, D. C.: A Tool for Content based Navigation of Music. In Proc. 6th ACM Int. Multimedia Conf., Bristol, UK (1998) 361–368

[3] Sonoda, T., Goto, M., and Muraoka, Y.: A WWW-based Melody Retrieval System. In Proc. Int. Computer Music Conf., Michigan, USA (1998) 349–352

[4] Kosugi, N., Nagata, H., and Nakanishi, T.: Query-by-Humming on Internet. In Proc. Int. Conf. on Database and Expert Systems Applications (2003) 589–600

[5] Tsuji, Y., Hoshi, M., and Ohmori, T.: Local Pattern of a Melody and Its Applications to Retrieval by Sensitivity Words. Technical Report of IEICE of Japan, Vol. SP96-124 (1997) 17–24

[6] Sato, A., Ogawa, J., and Kitakami, H.: An Impression-based Retrieval System of Music Collection. In Proc. of 4th Int. Conf. on Knowledge-Based Intelligent System and Allied Technologies, Brighton, U.K. (2000) 856–859

[7] Ikezoe, T., Kajikawa, Y., and Nomura, Y.: Music Database Retrieval System with Sensitivity Words Using Music Sensitivity Space. Trans. IPS of Japan, Vol. 42, No. 12 (2001) 3201–3212

[8] Omae, H., Ishibashi, N., Kiyoki, Y., and Anzai, Y.: An Automatic Metadata Creation Method for Music with Multiple Musical Instruments. Technical Report of IPS of Japan, Vol. 2001-DBS-125, No. 84 (2001) 145–152

[9] Kumamoto, T., and Ohta, K.: A Query by Musical Impression System using N-gram Based Features. In Proc. IEEE Conference on Cybernetics and Intelligent Systems (2004) 992–997

[10] Kumamoto, T.: Design and Implementation of Natural Language Interface for Impression-based Music-retrieval Systems. Lecture Notes in Artificial Intelligence, Vol. 3214, Springer-Verlag, Berlin Heidelbelg New York (2004) 139–147

[11] Kumamoto, T., and Ohta, K.: Design and Development of Natural Language Interface for an Impression-based Music Retrieval System. Technical Report of IPS of Japan, Vol. 2003-NL-153, No. 13 (2003) 97–104

[12] Harada, S., Itoh, Y., and Nakatani, H.: Interactive Image Retrieval by Natural Language. Optical Engineering, Vol. 36, No. 12 (1997) 3281–3287

[13] Ohno S., and Hamanishi, M.: Ruigo-Kokugo-Jiten. Kadokawa Shoten Publishing Co.,Ltd., Tokyo, Japan (1985)

[14] Kan, T.: Multivariate Statistical Analysis. Gendai-Sugakusha, Kyoto, Japan (1996)

[15] Kurohashi, S., and Nagao, M.: Manual for Japanese Morphological Analysis System JUMAN Version 3.61. http://pine.kuee.kyoto-u.ac.jp/nl-resource/juman.html (1999)

[16] Taniguchi, T.: Music and Affection. Kitaooji Syobo Press, Kyoto, Japan (1998)