International Journal of Fuzzy Logic and Intelligent Systems, vol. 12, no. 1, March 2012, pp. 94-100
http://dx.doi.org/10.5391/IJFIS.2012.12.1.94 pISSN 1598-2645 eISSN 2093-744X

Web Catchphrase Improve System Employing Onomatopoeia and Large-Scale *N*-gram Corpus

Hiroaki Yamane and Masafumi Hagiwara*

Department of Information and Computer Science, Keio University, Yokohama, 223-8522, Japan

Abstract

In this paper, we propose a system which improves text catchphrases on the web using onomatopoeia and the Japanese Google *N*-grams. Onomatopoeia is regarded as a fundamental tool in daily communication for people. The proposed system inserts an onomatopoetic word into plain text catchphrases. Being based on a large catchphrase encyclopedia, the proposed system evaluates each catchphrase's candidates considering the words, structure and usage of onomatopoeia. That is, candidates are selected whether they contain onomatopoeia and they use specific catchphrase grammatical structures. Subjective experiments show that inserted onomatopoeia is effective for making attractive catchphrases.

Key words : catchphrase, onomatopoeia, sound symbolism, Large-scale *N*-gram Corpus

1. Introduction

Most human knowledge, and most human communication, are represented and expressed using language. Especially, written languages are influential. For example, catchphrases play an important role in the field of advertising. Catchphrases need to be appealing to people in a short amount of time or sentences. Therefore, being different from other written languages, catchphrases attract people by expressing the uniqueness and merits of an object briefly and effectively with short sentences.

In the field of advertising, there are a number of researches on analyzing taglines, slogans and catchphrases [1][2]. There are also advertising slogan generators [3][4] in which the keywords are in a fixed line. As for the sentence generation, Banko et al. [5] proposed a system that produces headlines based on statistical translation.

Recently, online advertising is becoming popular worldwide [6][7] and people take care of the contents there more than ever. Since advertisement on the web has become very common, it seems that there is necessity for making catchphrases automatically. Some catchphrases on the web show the name of the target and inform the users of the name. However, such texts are sometimes too plain and not enough for conveying the sentiments to the users in many cases.

Here, we employ Japanese onomatopoeia to appeal and attract users on the point of sensitivity. Onomatopoeia is a group of imitative and echoic words, which is thought to be relatively fundamental compared to the other words because it is said that infants tend to grow with listening to onomatopoeia [8]. Moreover, from the field of neuroscience, onomatopoeia activates more in brain region suggesting that onomatopoeic sounds can serve as a bridge between language and nonverbal sound [9]. There are also some onomatopoeia-related researches in engineering such as applying its activity of sounds to motion of a robot [10], constructing an onomatopoeia map [11] and onomatopoeia hierarchical clustering [12].

This paper proposes a new system which generates improved web text catchphrases. In the proposed system, we employ Japanese Google *N*-grams [13] for obtaining the onomatopoeias in first generation. After the process, the proposed system computes sensitivities of additional onomatopoeias from the relationship between sensitivity words.

The rest of this paper is organized as follows: Section II proposes the web catchphrase improve system. Section III introduces the evaluation experiment and Section IV discusses the topic. Finally, Section V concludes the paper.

2. Web Catchphrase Improve System

Fig. 1 illustrates the overall flow of the proposed system. It consists of 4 parts.

First, web catchphrases are fed to the proposed system as inputs. After that, using Google *N*-gram search engine [14], the proposed system generates catchphrase candidates. Third, candidates are scored based on grammatical

Manuscript received Feb. 28, 2012; revised Mar. 14, 2012; accepted Mar. 19, 2012.
*Corresponding Author(hagiwara@soft.ics.keio.ac.jp)
It is the extended version of a paper awarded as a best presentation in ISIS2011.

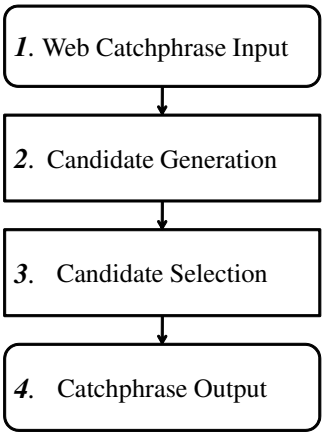

Figure 1: The Overall Flow of the Proposed System

structures and usage of onomatopoeia. Finally, after sorting candidates by the scores, the proposed system outputs selected ones as the final ones.

2.1 Web Catchphrase Input

The proposed system is intended to make text catchphrases on the web more elaborate. Such kinds of the catchphrases are fed to the system.

2.2 Candidate Generation

Here, we will describe how the proposed system generates catchphrase candidates.

2.2.1 Requesting Candidates

To obtain various kinds of sentence information, the Japanese Google *N*-gram Corpus [13] is employed. It contains 1 to 7grams with frequencies, which are constructed from 20 billion sentences. Since Google *N*-gram Corpus contains extremely enormous numbers of words, ways of searching candidates are important. We have used the Search System for Giga-scale *N*-gram Corpus named *ssgnc* [14] for a quick search. The system accepts several input words and outputs candidates with frequency. The proposed system requests an onomatopoetic word between two combination words in the web text catchphrases using *ssgnc*.

2.2.2 Onomatopoeia Existence Check

The proposed system checks the existence of onomatopoeia using onomatopoeia dictionary [15], which contains about 4,500 onomatopoetic words. If candidates contain onomatopoeia, they will proceed to the next step.

2.3 Candidate Selection

To select attractive catchphrases, we employ two kinds of elimination strategies. The first one is grammatical structure and the last one is onomatopoeia sensitivity.

2.3.1 Grammatical Structure

To select more suitable catchphrases, there is necessity of checking the structure of the sentences. Before constructing the proposed system, we have analyzed the uniqueness of catchphrases using the catchphrase corpus [16]. The book lists more than 6,400 catchphrases and covers 12 fields such as food and livingware. By using the Japanese morphological analyzer called MeCab [17][18], we have gathered the data concerning *N*-gram parts of speeches (POS) tagging as knowledge of grammatical structure of catchphrases.

In catchphrases, the more the grammatical structure appears, the more it can be regarded as suitable structure for catchphrases. On this point, two aspects of features of catchphrases – catchphrase corpus itself and multiple corpora – are considered.

2.3.1-i Grammatical structure score in catchphrase corpus

Since catchphrases have distinct grammatical structure, it seems that the more the structure of candidate is similar to the catchphrases in the corpus, the more suitable the candidates are. Using a probabilistic model, the score is calculated. The probability $P_c(L)$ that a POS sequence L appears in catchphrase corpus is given as:

$$P_c(L) = \frac{F(L)}{N} \tag{1}$$

where $F(L)$ is frequency of the POS sequence and N is the number of catchphrases in the corpus.

Assigning the Eq. (1) to definitional identity of information volume and summing the *N*-gram from 1 to 7, the total score S_{I} is calculated as:

$$S_{\mathrm{I}} = -\sum_{k=1}^{N} \log \frac{F(L)}{N}. \tag{2}$$

This processing is operated from both forward and backward.

2.3.1-ii Grammatical structure score using multiple corpora

Since catchphrases are different in terms of grammatical structure, there is necessity of selecting "catchphrase like" structure preferentially. Hence, the probability that POS sequence L is in the catchphrase corpus should be calculated. Due to the different number of sentences in each

corpus, we consider the probability $P_v(L)$ on frequency of normalized POS sequence. The formula is given as:

$$P_v(L) = \frac{P(L|C_k)}{\sum_C P(L|C)} \tag{3}$$

where each corpus indicates C and especially the targeted catchphrase corpus is C_k. Assigning catchphrase, encyclopedia [19], blog [20] and dialogue corpus [21] to Eq. (3), an expression

$$\frac{P(L|C_k)}{\sum_C P(L|C)} \approx \frac{\frac{F_{C_k}(L)}{N_{C_k}}}{\sum_{C_x \in corpora} \frac{F_{C_x}(L)}{N_{C_x}}} \tag{4}$$

is obtained. Therefore, computing *N*-gram from 1 to 7, the score S_{II} indicating grammatical structure one using multiple corpora is calculated as:

$$S_{\mathrm{II}} = -\sum_{k=1}^{N} \log \frac{\frac{F_{C_k}(L)}{N_{C_k}}}{\sum_{C_x \in corpora} \frac{F_{C_x}(L)}{N_{C_x}}}. \tag{5}$$

In the case that there is no corpus containing a grammatical structure, it is computed as 1 occurrence.

As well as single catchphrase grammatical computation, the proposed system calculates in terms of both forward and backward.

After calculating different types of scores, to select more catchphrase-like candidates, the catchphrase candidates whose amount of information are less will remain as candidates.

2.3.2 Onomatopoeia Suitability

Onomatopoeia suitability is estimated by the product of onomatopoeia sensitivities and their sensitivity words' relationship to the catchphrase candidate.

First, we describe the χ^2 for calculating the word relation. Pointwise Mutual Information (PMI) has been used for estimating the related terms. However, when there is variance of frequencies, it becomes more difficult to obtain a good result. Therefore, we employed χ^2 values [22] for extracting related terms. This computation limits influence on variance of occurrence probabilities because χ^2 values are calculated by normalized term frequencies from word group.

χ^2 is given as follows:

$$\chi^2(w_i, w_j) = \frac{n(w_i, w_j) - E(w_i, w_j)}{E(w_i, w_j)} \tag{6}$$

where $n(w_i, w_j)$ is actual co-occurrence of word w_i and w_j,

Table 1: Phonemes with Sensitivities. Note that Cons. indicates consonant. Hd (Hard), St (Strong), Hm (Humid), Sm (Smooth), Rd (Round), Ec (Elastic), Sd (Speedy), Wm (Warm)

	Hd	St	Hm	Sm	Rd	Ec	Sd	Wm
Vowel								
A	0	1	-1	1	2	-2	-1	0
I	2	2	0	0	-1	1	2	-1
U	-1	-1	2	0	2	2	0	2
E	1	-2	2	0	-2	0	0	2
O	-1	2	0	1	2	0	-2	1
Cons.								
K	2	2	1	0	0	0	2	-1
S	2	0	1	2	0	0	2	-1
T	2	1	2	2	0	1	-1	-2
N	-1	0	2	-1	1	0	-2	2
H	-2	-2	1	0	1	-1	-1	2
M	-2	-2	1	0	2	0	-1	2
Y	-2	-1	0	1	2	1	0	0
R	-1	-1	2	1	0	2	1	0
W	-2	2	1	0	2	0	0	1
Others								
d (etc)	1	1	-1	-1	-1	0	-1	0
p	-1	-1	0	0	1	1	1	1
with y	-1	-1	1	0	1	2	2	1
t	0	0	0	0	0	0	1	0

$$E(w_i, w_j) = S_{wi} \times \frac{S_{wj}}{S_G} \tag{7}$$

where $S_{wi} = \sum_k n(w_i, w_k)$ is the summation of co-occurrence between word w_i and word group G. $S_G = \sum_{w_i \in G} S_{wi}$ is given as the summation of co-occurrence of all words.

Second, the system calculates sensitivity. We used Onomatopoeia Dictionary [15] for obtaining large number of onomatopoeias. Komatsu et.al [10] have proposed a new system to show the movement image of onomatopoeias. They use 8 dimension vectors for the system. Table 1 shows phonemes with sensitivities.

Two letters each on the top show abbreviated name. In detail, the each indicates Hd (Hard), St (Strong), Hm (Humid), Sm (Smooth), Rd (Round), Ec (Elastic), Sd (Speedy), Wm (Warm) respectively. Komatsu et.al also defined an equation from phonemes sensitivities to onomatopoeia one. To apply XYXY (Ex. Waku-Waku: State of Excitement) type onomatopoeia, the system computes the vectorized sensitivities O_i as:

Web Text Catchphrases

1. 価格com の中古車検索
   Used car retrieval on Kakaku com
2. 塩素を取っちゃって良いんですか
   May I take chlorine?
3. トリプル美白キャンペーン
   Triple beautification and whitening campaign
4. 神奈川リフォーム業者探し
   Search for Kanagawa reform supplier
5. 絹スパッツ・レギンス通販
   Silk spats leggings mail order

Figure 2: Some Examples of the Web Text Catchphrases

$$O_i = 2xc_i + xv_i + \frac{2yc_i + yv_i}{2} \tag{8}$$

where xc_i is the *i* th consonant value in X, xv_i is the *i* th vowel value in Y, yc_i is *i* th consonant value in X and xv_i is the *i* th vowel value in Y.

We employ both Eq. (8) and Eq. (6) for obtaining the sentence sensitivity. In Eq. (6), the proposed system computes the χ^2 values between words in candidates and sensitivity words. The onomatopoeia similarity score S is given as:

$$S_o = \sum_{i=1}^{8} (O_i \times S_{i\chi^2}) \tag{9}$$

where $S_{i\chi^2}$ is the summation of χ^2 values related to sensitivity word.

2.4 Catchphrase Output

After computing those given scores, the system outputs the sorted candidates as the final one.

3. Experiments

We conducted subjective experiments described below to evaluate the performance of the proposed system. After defining the condition we show the result of experiments.

We have used text advertisement catchphrase on gmail.

3.1 Conditions

In order to evaluate the quality of the generated catchphrases, we carried out *Turing test* like experiments. We shuffled 20 system-generated ones and 20 catchphrases on the web catchphrase. Both five examples are shown in Fig. 2 and in Fig. 3 respectively.

Generated Candidates

1. ORBISの肌がツルツルと潤う化粧水/新発売
   Lotion / new sale to be enriched if skin of ORBIS is soft and smooth.
2. ドキドキとシクシクが凄い恋愛ゲーム
   The love game that makes you throb and great in sorrow.
3. なに？このシトシト泡。すっごい
   What? This steady bubbles. Amazing.
4. 車中泊でキチキチとドライブ旅行
   Sleep in a car and have a drive trip in tight schedule.
5. 高品質のキチキチ音を防止、防音シートです。
   It is prevention, and a sound-proof sheet as for the [kichikichi] sound of the high quality.

Figure 3: The Output Examples of the Proposed System. Note that underlines indicate onomatopoetic word.

Five subjects evaluated 40 catchphrases from three view points: The degree of correctness in sentences ; attracting attention level ; funny (interesting) level, without knowing which is the system-improved one.

We set the evaluation items as follows.

1. Is it a correct sentence?

Yes	———	3
Partially yes	———	2
No	———	1

2. Does it attract attention?

Yes	———	3
Partially yes	———	2
No	———	1

3. Which is the funny (interesting) level of the catchphrase?

High	———	5
Somewhat High	———	4
Average	———	3
Somewhat Low	———	2
Low	———	1

3.2 Result

We show the result of comparisons one by one. First, the result of validity of both the system-generated and the original catchphrases is illustrated in Fig. 4. Also, the score of difference is represented in Table 2.

As the table illustrates, the variance of the proposed system generated catchphrases is less than that of catchphrase encyclopedia. *F*-test did not prove unequal variance and *T*-test guaranteed the result 5% significance level ($p = 1.66 \times 10^{-6} < 0.05$). Second, the result of attractiveness of both the system-generated and the original catchphrases is illustrated in Fig. 5. Also, the score of difference is represented in Table 3.

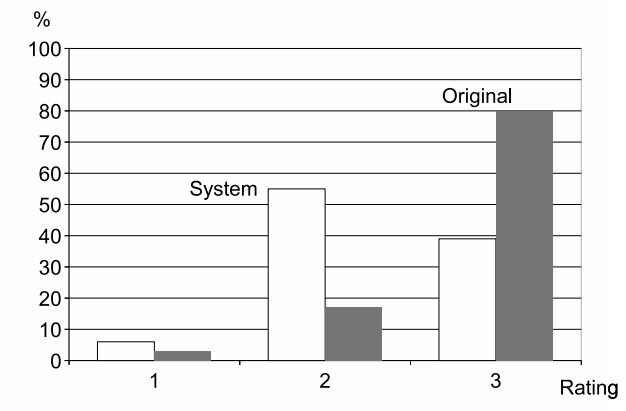

Figure 4: Appropriateness of both system-generated and original catchphrases

Table 2: Comparison of validity in terms of average and variance

	System-generated	Original
Average	2.33	2.77
Variance	0.0727	0.102

As the table illustrates, the average of the proposed system generated catchphrases is bigger than that of original ones. *F*-test showed homoscedasticity and *T*-test did not guaranteed the result 5% significance level ($p = 0.17 > 0.05$).

Finally, the result of funny level of both the system-generated and the original catchphrases is illustrated in Fig. 6. From Fig. 6 system fixed catchphrases outweighed the number in rating 3 to 5. As for rating 1, the proposed system seems to able to succeed in generating funny (interesting) catchphrases. *F*-test did not prove unequal variance

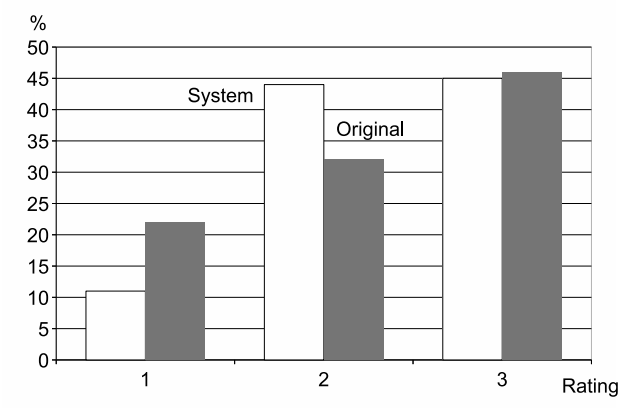

Figure 5: Attractiveness of both system-generated and original catchphrases

Table 3: Comparison of specificity in terms of average and variance

	System-generated	Original
Average	2.34	2.24
Variance	0.0089	0.125

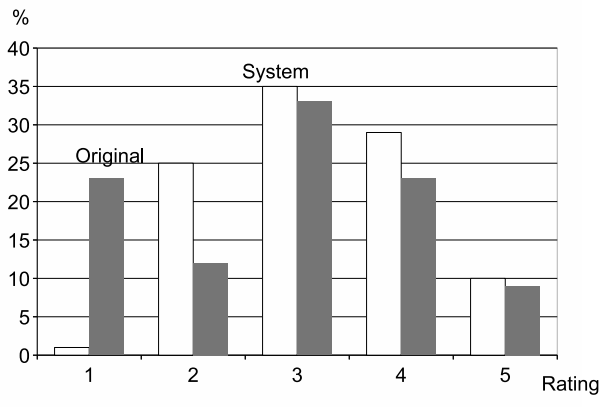

Figure 6: Funny level of both system-generated and original catchphrases

and *T*-test showed the funny level of the items generated by the proposed system outweighed that of the original in 5% significance level ($p = 9.9 \times 10^{-3} < 0.05$). Also, the score of difference is represented in Table 3.

As the table illustrates, the variance of the proposed system generated catchphrases is less than that of catchphrase encyclopedia.

4. Discussion

Due to the employ of onomatopoetic words, the proposed system is able to improve plain text catchphrases in terms of funny level. Especially as shown in Fig. 6, it seems that the proposed system contributes to reduce the number of boring plain catchphrases. However, system-generated ones lacked sentence appropriateness. Probably, this is because the proposed system is not able to consider the meaning of catchphrases. Thus, to make the system more applicable in general usage in advertisement, we

Table 4: Comparison of specificity in terms of average and variance

	System-generated	Attract
Average	3.22	2.83
Variance	0.189	0.325

think the integration of meaning and sensitivity is necessary.

5. Conclusion

In this paper, we proposed a system which improves text catchphrases on the web using onomatopoeia and the Japanese Google *N*-grams. Onomatopoeia is regarded as a useful tool in daily communication for human beings. The proposed system inserts an onomatopoetic word into plain text catchphrases. Being based on a large catchphrase encyclopedia, the proposed system evaluates each catchphrase's candidates considering the words, structure and usage of onomatopoeia. That is, candidates are selected whether they contain onomatopoeia and they use specific catchphrase grammatical structures. Subjective experiments show that inserted onomatopoeia is effective for making attractive catchphrases.

Our future work is to make it in more general way such as expanding the idea of sound symbolism to normal sentences perceptions.

References

[1] C. Kohli, L. Leuthesser, and R. Suri, "Got slogan? guidelines for creating effective slogans," *Business Horizons*, vol. 50, no. 5, pp. 415–422, 2007.

[2] H. Kitamura, R. Yamaji, and H. Tabuki, *Adverting Catchphrase*. Yuhikaku, 1981.

[3] Sloganizer.net, "Instant slogans with our slogan generator." http://www.sloganizer.net/en/.

[4] THE-PCMAN-WEBSITE, "Free slogan generator." http://www.thepcmanwebsite.com/media/free_slogan_generator/index.php.

[5] M. Banko, V. O. Mittal, and M. J. Witbrock, "Headline generation based on statistical translation," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pp. 318–325, 2000.

[6] "Online advertising: A 59 billion-euro market in 2012, up from 31 billion in 2008." http://www.international-television.org/archive/online-advertising-world-usa-europe_2005-2012.pdf.

[7] "Worldwide internet advertising spending to surpass $106 billion in 2011." http://www.marketingcharts.com/television/worldwide-internet-advertising-spending-to-surpass-106-billion-in-2011-5068/.

[8] C. Kit, "How does lexical acquisition begin? A cognitive perspective," *Cognitive Science*, vol. 1, no. 1, pp. 1 – 50, 2003.

[9] T. Hashimoto, N. Usui, M. Taira, I. Nose, T. Haji, and S. Kojima, "The neural mechanism associated with the processing of onomatopoeic sounds," *NeuroImage*, vol. 31, no. 4, pp. 1762 – 1770, 2006.

[10] T. Komatsu and H. Akiyama, "Expression system of onomatopoeias for assisting users' intuitive expressions," *The Transactions of the Institute of Electronics, Information and Communication Engineers. A*, vol. 92, no. 11, pp. 752–763, 2009.

[11] Y. Tomoto, T. Nakamura, M. Kanoh, and T. Komatsu, "Visualization of similarity relationships by onomatopoeia thesaurus map," in *Fuzzy Systems (FUZZ), 2010 IEEE International Conference on*, pp. 1–6, 2010.

[12] K. Komiya and Y. Kotani, "Classification of Japanese onomatopoeias using hierarchical clustering depending on contexts," in *Computer Science and Software Engineering (JCSSE), 2011 Eighth International Joint Conference on*, pp. 108–113, 2011.

[13] T. Kudo and H. Kazawa, "Web Japanese N-gram version 1." Gengo Shigen Kyokai, 2007.

[14] S. Yata, "Search system for giga-scale ngram corpus." http://code.google.com/p/ssgnc/.

[15] M. Ono, *Japanese Onomatopoeia Dictionary: echoic and imitative words 4500*. Shogakukan, 2007.

[16] Y. Kuno, *Catalog and Flier Catchphrase Encyclopedia*. PIE BOOKS, 2008.

[17] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," in *In Proc. of EMNLP*, pp. 230–237, 2004.

[18] T. Kudo, "Mecab: Yet another part-of-speech and morphological analyzer." http://mecab.sourceforge.net/.

[19] "Wikipedia Japanese archive." http://dumps.wikimedia.org/jawiki/.

[20] Kurohashi Laboratory: Graduate School of Informatics in Kyoto University, "KNB Corpus (Kyoto-University and NTT Blog Corpus)." http://nlp.kuee.kyoto-u.ac.jp/kuntt/, 2009.

[21] Uemura Laboratory: Faculty of Environmental Engineering in the University of Kitakyushu, "Hypermedia Corpus of Spoken Japanese." http://www.env.kitakyu-u.ac.jp/corpus/texts/index.html, 1996.

[22] T. Sakaki, Y. Mtsuo, K. Uchiyama, and M. Ishizuka, "Construction of related terms thesauri from the web," *Journal of natural language processing*, vol. 14, no. 2, pp. 3–31, 2007.

Hiroaki Yamane
Hiroaki Yamane received his B.E. and M.E. in information and computer science from Keio University, Yokohama, Japan, in 2010 and 2011. Currently, he is in the doctoral course at Keio University. His research interest focuses on integrating natural language processing and knowledge from brain science.
E-mail : yamane@soft.ics.keio.ac.jp

Masafumi Hagiwara
Masafumi Hagiwara received his B.E., M.E. and Ph.D degrees in electrical engineering from Keio University, Yokohama, Japan, in 1982, 1984 and 1987, respectively. Since 1987 he has been with Keio University, where he is now a Professor. From 1991 to 1993, he was a visiting scholar at Stanford University. He received the Niwa Memorial Award, Shinohara Memorial Young Engineer Award, IEEE Consumer Electronics Society Chester Sall Award, Ando Memorial Award, Author Award and Contribution Award from the Japan Society of Fuzzy Theory and Systems (SOFT), Technical Award and Paper Award from Japan Society of Kansei Engineering. His research interests include neural networks, fuzzy systems, evolutionary computation and kansei engineering. Dr. Hagiwara is a member of IEICE, IPSJ, SOFT, IEE of Japan, Japan Society of Kansei Engineering, JNNS and IEEE (Senior member).
E-mail : hagiwara@soft.ics.keio.ac.jp