PAPER

Sound database retrieved by sound

Sample of SUB-TITLE

Hai Qi*, Pitoyo Hartono†, Kenji Suzuki‡ and Shuji Hashimoto§

Department of Applied Physics, Waseda University, Ohkubo 3–4–1, Shinjuku-ku, Tokyo, 169–8555 Japan

(Received June 22 2001, Accepted for publication June 11 2002)

Abstract: In this paper, we propose a sound-database system, which is able to extract stored data using sound as a key for the query. This ability realizes the sound extraction without having to specify the acoustical characteristics of the sound. The system repetitively searches and presents sounds, which have similarity in timbre to the key sound, until the user finds a satisfactory sample. The parameters that characterize a sound's timbre, which is a psychoacoustical factor for sound perception, are adopted as the sound's indices in the database and used for similarity matching in the searching process. Because the definition of similarity in sounds differs from user to user, the proposed system is equipped with an adaptive preference-weighted searching mechanism that adapts its searching focus based on the user's preference. Because of the ability of the proposed system to realize an intuitive query, this system can be broadly used by a user without special acoustical knowledge.

Keywords: Sound database, Adaptive query, Intuitive data retrieval

PACS number: 43.60.Lq, 43.66.Jh

1. Introduction

In this paper, we propose a sound database system, which can implement an intuitive sound query using sound as the key for the query. The query process for the proposed system searches the database for sounds that have similarities in timbre with the sound given as the key; consequently in the proposed system, a sound in the database is indexed by a number of parameters that characterize the timbre. Because of the difference in the sound's preference from one person to another, the proposed system is equipped with an adaptive preference-weighted query mechanism, which adapts the searching focus according to the preference of the user. Until now, in searching for data, users have had to use objective features such as physical or logical characteristics of sound that can be difficult or even stressful. The proposed method realized a data query system with intuitive impressions of the data rather than acoustical or physical features, which may increase the database usability. For example, consider a person creating sound effects for background music for a song or other purposes. The person may have an ideal sound in mind but may not be able to pinpoint the sound on the basis of its acoustical or physical features. It would be easier to mimic the sound and search for similar sounds in the database. The extracted sounds might not perfectly match the person's desired one, but they would be close to it, thus not difficult to modify. In the past, a number of sound database systems have been built [1]. Some database systems adopted a textual information system for indexing the samples in the database. Query in the database systems is done by matching the text information between the query-key and the members of the database. This kind of database system is different from the proposed system, because these systems required the users to know about the rules for data indexing, while in the proposed system the user can intuitively search the database without prerequisite knowledge. These systems also required the database designers to assign specific textual information to each sample in the database, which can be very time consuming, while in the proposed database system, the timbre parameters that are used as indices are automatically extracted.

Blum [2] proposed a sound database system that utilizes sound as the query-key. This system differs from the proposed system, because it adopted acoustical attributes for indexing, implying that it requires the user to have knowledge about the relation of the acoustical attributes to the perceptual sound attributes.

* e-mail: qihai@shalab.phys.waseda.ac.jp
† e-mail: hartono@shalab.phys.waseda.ac.jp
‡ e-mail: kenji@shalab.phys.waseda.ac.jp
§ e-mail: shuji@shalab.phys.waseda.ac.jp

Keislar [3] and Vertegaal [4] proposed database systems that enabled the users to add new samples to the database and also modify the already existing data to form new ones. Feiten [5] utilized a neural network to generate a mapping rule between the sound's acoustical attributes and its perceptual attributes, but once the rule was fixed, it could not accommodate different users' preferences. Our main focuses in constructing the proposed database systems are:

(1) Automatic indexing of samples to be added to the database.

(2) Realizing a content-based retrieval without requiring the users to have prerequisite acoustical knowledge.

(3) Accommodating different users' preferences in data searching.

The three features mentioned above could be dealt with by building a database system that automatically extracts the sound's timbre parameters and utilizes them for data indexing. The query is based on the similarity of the query-key's timbre parameters with the existing data's timbre parameters. The third focus mentioned above can be dealt with by assigning weight to each of the parameters to indicate their importance in the data searching process. Different users may assign different weight configurations for similarity descriptions, so the weights will be updated for every searching process. In this paper, the proposed sound database is explained in Section 2. The experimental results are given in Section 3, and the conclusions will be given in the final section.

2. Sound database

A block diagram of the proposed system is illustrated in Fig. 1. In the data storage process, the system automatically extracts the new sound's timbre parameters and assigns them as the sound's index, then stores the sound in the sound database and the related index in the parameter database. In the query process, the system extracts the timbre parameters from the sound given as the query-key. Based on the timbre parameters and the preference of the user, represented by weight values that give indication of the importance of each parameter according to the user, the system searches the database to extract data that have similarity with the key, and return the searching results in the form of stereo sounds.

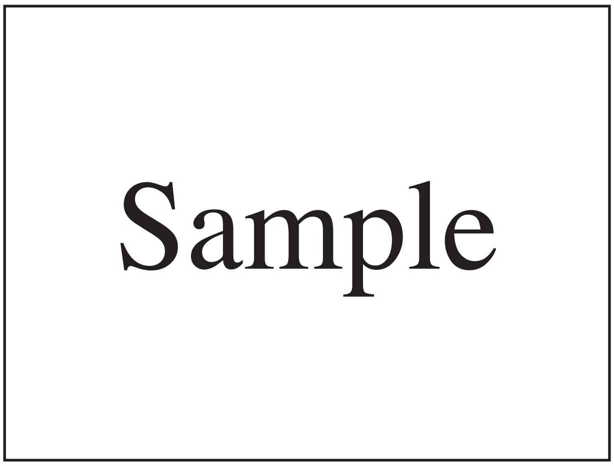

Fig. 1 System overview.

2.1. Sound Data

The sound data are stored at the sampling frequency of 44.1 kHz and a quantizing level of 16 bits with WAVE format. There are currently data of approximately 1,800 sounds in the system, with a total size of 200 MB). These sounds were recorded or selected from various sources, such as CDs of musical instruments, sound effects, and sounds gathered via the Internet. Before being added to the database, adjustment was made to each sound by removing the silent parts and normalizing the amplitude.

Unlike conventional database systems, the data in the proposed system are not classified. Although classifying data and building hierarchical structures based on the classification methods will improve the query time and the cost of data management, these methods have a drawback in that the data have to be arranged according to the designer's preference and interpretation. This implies that once the database is built, it will lose its flexibility, so that the users will be forced to adapt to the rules set by the database's designer. The proposed system avoids this problem by deliberately not interpreting the data, so that the users can access them according to their personal preferences

2.2. Sound Attributes

A great deal of physical parameters can be extracted from sounds. In building the proposed system, it is very important to determine the relation of a sound's perceptual impression and its physical parameters. In this study, we consider that this perceptual impression is best expressed as "timbre."

Numerous studies have been done to investigate the relation between the physical properties of a sound and its position in the "perceptual space." Middo [6,7] conducted a series of studies about a sound's tone (impression) using the Semantic Differential Method. They confirmed that there are roughly three components that form the impression of a sound, which are "PLEASANT," "METALLIC" and "POWERFUL."

Research on "auditory scene analysis," which is a

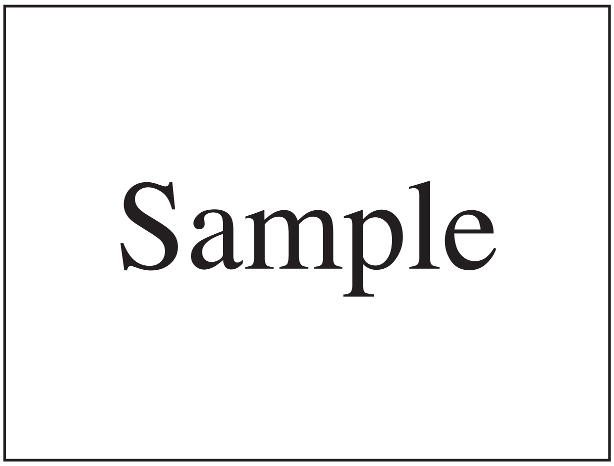

Fig. 2 Relations among timbre, loudness, pitch, and subjective duration.

function to relate the auditory physical stimuli and its psychological influences, has been done [8–11]. Taking into account the result of past research, in this research we concentrated our attention on the relation between a sound's timbre, which can be considered one of the psycoacoustical factors, and the sound's physical parameters. According to psychological studies, there are three major attributes that are used by humans to distinguish sounds. Two important perceptual attributes are "loudness" and "pitch." The other one is timbre. Timbre is defined as 'an attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar' (American Standards Association, 1960). However, many researchers have argued that "timbre" is not independent of "loudness," "pitch" and another attribute called "subjective duration." The relations between these four attributes are shown in Fig. 2. Based on these relations, we can say that loudness, pitch and subjective duration have significant effect on the timbre.

2.2.1. Envelope feature

Experiments by Namba [12] and Miller [13] confirmed that there is a strong correlation between the durations of attack and sustain of a sound's wave with its timbre. To extract the features from the original sound wave (Fig. 3(a)), the sound is simplified as shown in Fig. 3(b). The extracted features are the attack duration (t_1), decay duration (t_2), sustained duration (t_3), release duration (t_4) and the ratio of the sustained level (r_1), which can be calculated according to:

$$r_1 = \frac{SL}{ML} \tag{1}$$

2.2.2. Spectral feature

The physical characteristics of the sustained period of a sound make a great contribution to its timbre [14]. However, Stevens [15] and Namba [12] pointed out that the physical characteristics of the attack period of a sound also play an important role in influencing its timbre. So we extracted parameters from the spectra of the attack and sustained periods. The spectral parameters are extracted by Cepstrum Analysis. At first, as the most important parameter, the fundamental frequencies from the attack period (f_{a}) and the sustained peperiod (f_{s}) are extracted. The difference between the fundamental frequencies is also calculated and used as a parameter. From each spectral envelope, a factor that is defined as "harmonicity," which is the sum of the power of fundamental frequency and all of the harmonic frequencies relative to the sum of the all of the powers, can then be calculated. This parameter indicates the proportion of harmonic components in the sound, which will be significant for generating the impression

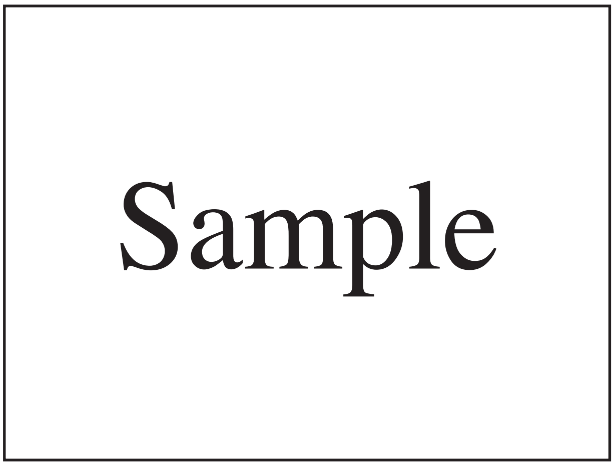

Fig. 3 Sound's envelope.

of "PLEASANT." The "harmonicities" for the attack and sustained periods are expressed as r_2 and r_3, respectively. The spectral envelope is divided into 5 octave bands (0–1.5 kHz, 1.5–3 kHz, 3–6 kHz, 6–12 kHz, 12 kHz) and the ratio of the average powers of the 5 octave bands to the average power of all partials is extracted. They are represented by $ba_1 - ba_5$ and $bs_1 - bs_5$, respectively. This idea is based on the following research.

(1) Plomp [16] and de Bruijn [17] confirmed that, for two sounds, the physical distance between the spectrum envelopes corresponds to the distance in psychological space.

(2) It is known that the power spectra of the human voice are focused on a frequency band between 150 Hz and 6 kHz called the "speech band." So we set 6 kHz as a boundary. The formant frequency can help us to recognize the phoneme. (Stevens [15] and other research [18–20] confirmed that the formant character can also be found in some musical tones). For example, the first formant frequency of five Japanese vowels are in between 300–1,200 Hz, but the second formant frequency of /i/ /e/ are in between 1,800–3,000 Hz, that of /u/ /o/ /a/ are between 900–1,500 kHz and the anti-formant frequency of the nasal consonant [m] is in between 500–1,500 Hz, that of the nasal consonant [n] is in between 2.0–3.0 kHz, and so on. Thus, we also set 1,500 Hz and 3,000 Hz as boundaries.

(3) In the proposed system, the sound data are stored at the sampling frequency of 44.1 kHz, but because the sounds were collected from various sources, some sounds were sampled at the sampling frequency of 22.02 kHz or 24 kHz. We cannot extract the frequency components higher than 12 kHz from these sounds; therefore, we set 12 kHz as a boundary.

(4) Kondo [21] conducted an experiment to find the relation between the dips in the frequency characteristics of the sound reproducing system and its sound quality and evaluated this using preference tests. The results of the experiment explained the relationship between the impression of "PLEASANT" and the center frequency of the dips. At the points of 600 Hz, 2,400 Hz, 3,600 Hz, and 7,200 Hz, two groups of observers who have quite different preference patterns made quite different judgments regarding the degree of PLEASANT. Because the four points are exactly in our four octave bands, this implies that different users' preferences can be accommodated in the searching process by giving different weights for parameters corresponding to these bands.

2.2.3. Harmonic features

In the proposed database systems, there are about 200 musical tones generated by musical instruments. To distinguish these musical tones efficiently, some harmonic parameters have to be taken into account. The most important feature of a musical tone is the acoustical spectrum. In the proposed system, the ratio of the powers of 20 harmonic frequencies to the power of the fundamental frequency from the each spectrum (attack and sustain) are calculated and represented by $ha_1 - ha_{20}$ and $hs_1 - hs_{20}$, respectively.

Saito [22] pointed out that the varieties of the rise and fall of each harmonic can play an important role in the timbre of a musical tone. The proposed system also refers to the synchronicity of each harmonic envelope and extracts the differences in the temporal points when the harmonics reach the onsets, the maximum point, the release point (50% of maximum level), and the offsets. Figure 4 illustrates the differences in each point of all harmonics. The following parameters are adopted as indices, where ON_i, MAX_i, REL_i, OF_i represent the onset, maximum, release and offset amplitudes of the i-th harmonic, respectively.

$$\begin{aligned} d_1 &= \sqrt{\sum_{i=2}^{20}(ON_i - ON_0)^2}, \\ d_2 &= \sqrt{\sum_{i=2}^{20}(MAX_i - MAX_0)^2} \\ d_3 &= \sqrt{\sum_{i=2}^{20}(REL_i - REL_0)^2}, \\ d_4 &= \sqrt{\sum_{i=2}^{20}(OF_i - OF_0)^2} \end{aligned} \tag{2}$$

2.3. Search Mechanism

The data searching procedure is shown in Fig. 5. The proposed system executes an iterative searching mechanism, in which each query yields a number of sounds, which have the greatest similarities with the query-key. The user selects one sound that best matches the user's preference. Based on the choice, the system renews the preference weights and use the chosen sound as the new key. By renewing the preference weights, the system adaptively accommodates the preference of the user to speed up the searching process. This process is iterated until the user finds a satisfactory sound.

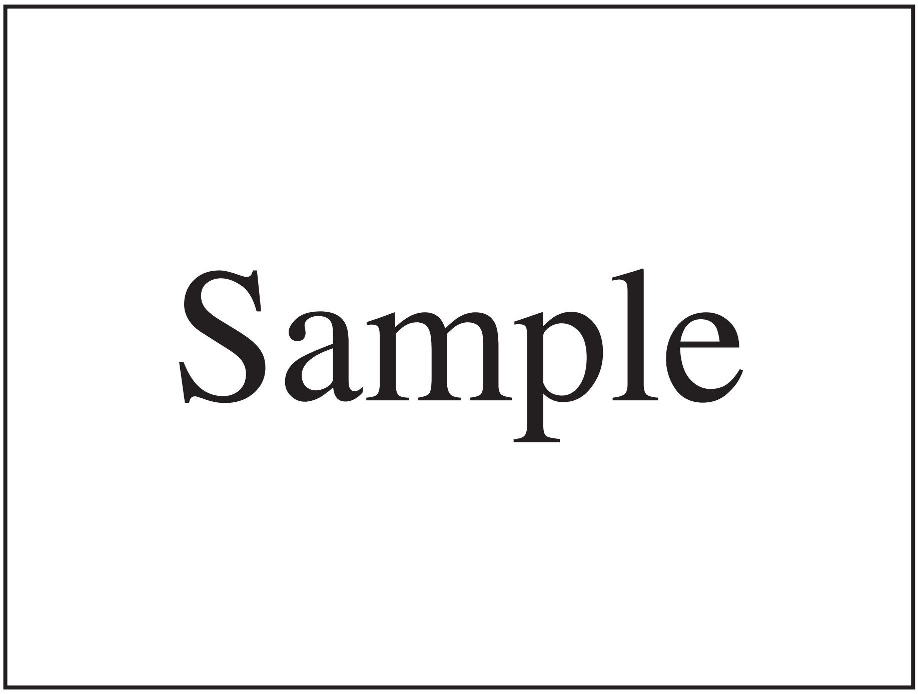

Fig. 4 Harmonic parameters.

Table 1 Extracted parameters.

Envelope features (5)	$t_1 - t_4, r_1$
Spectral features (15)	attack: $f_{\rm a}$, $ba_1 - ba_5$, r_2
	sustain: $f_{\rm s}$, $bs_1 - bs_5$, r_3, $f_{\rm d} = f_{\rm a} - f_{\rm s}$
Harmonic features (44)	attack: $ha_1 - ha_{20}$
	sustain: $hs_1 - hs_{20}$, $d_1 - d_4$

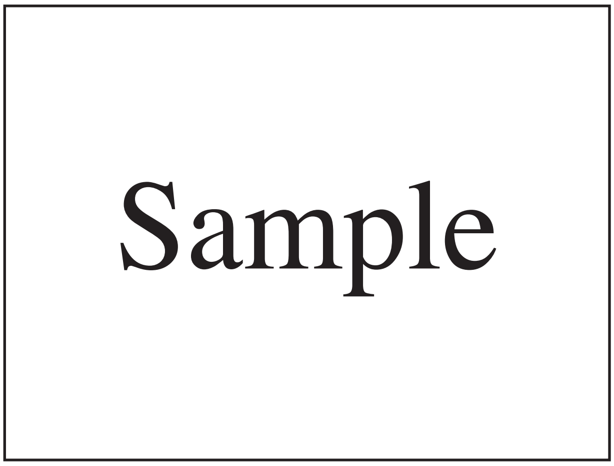

Fig. 5 Data searching process.

2.3.1. Index matching

After a sound was given as a query-key, the system extracts timbre parameters from the sound as shown in Table. 1. The system will then execute index matching with all the data in the database according to the following

$$
\begin{aligned}
E_k &= \sum_{j=1}^{64} a_j \frac{(X_j - {X_j}^k)^2}{\kappa_j} \\
\kappa_j &= X_j \quad \text{for} \quad X_j \quad (X_j > 0.1) \\
&= 0.1 \quad \text{for} \quad X_j \quad (X_j \leq 0.1)
\end{aligned} \qquad (3)
$$

where E_k denotes the distance between the key sound and the k-th sound in the database. X_j and ${X_j}^k$ show the j-th parameters of the key sound and the k-th sound in the database, respectively. The weight of the j-th parameter is denoted by a_j.

2.3.2. Weight adaptation

The idea of the weights renewal executed for every search iteration is illustrated in Fig. 6. For the purpose of simplicity, in Fig. 6, it is shown that each sound is indexed by 2 timbre parameters, which are x and y. In this example, 5 sounds are generated by the system after the key sound is given. Suppose that the user chooses Sound 2 as the sound most similar to the key sound. From Fig.

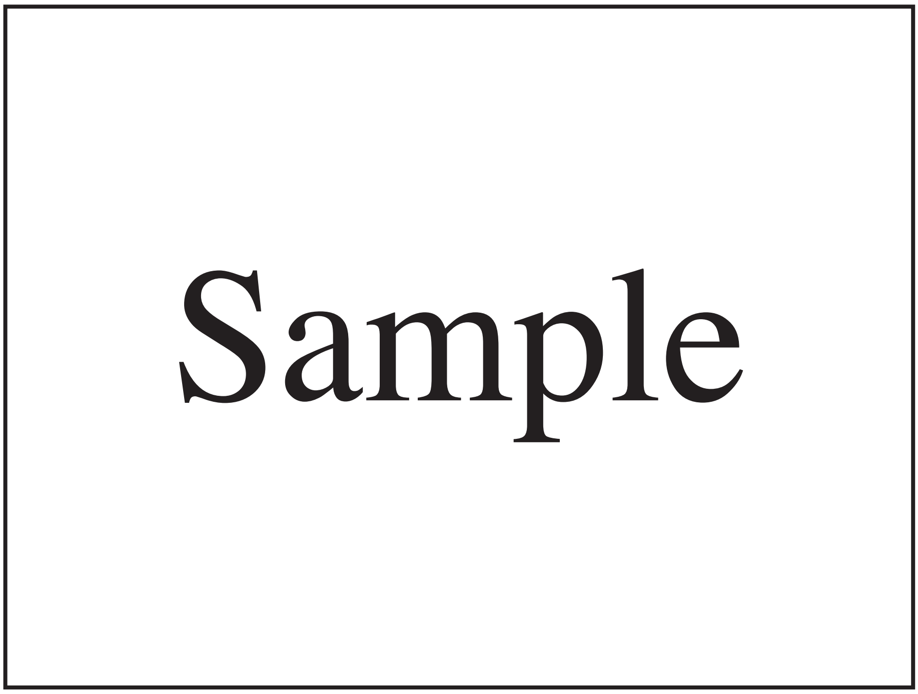

Fig. 6 Feature parameter space.

6(a), it is clear that Sound 2 has the largest difference relative to the key sound in terms of the parameter y, but because the user chose Sound 2, it should be closer to the key sound in the perceptual space of the user. The interpretation of the system with regard to the choice of the user is that, for the given sound, the importance of the parameter y is low, so that the related weight should be decreased. The weights renewal will be executed until the distance of the chosen sound from the key sound (Eq. (3)) becomes the smallest. Changing the preference weights is equivalent to changing the scale of each parameter, so that eventually the query is executed inside the user's preference space. The weight renewal is executed as follows: First, calculate the difference between the j-th index of the key sound and all of the candidate sounds, as follows,

$$D_j{}^i(t) = |X_j{}^k(t) - X_j{}^i(t)| \tag{4}$$

in which $D_j{}^i(t)$ is the difference between the j-th index of the key sound and the i-th candidate sound at the t-th search. The distances are then ranked in small to large order. When the rank of the chosen sound is $R^s(t)$, the weight correction is then done as follows,

$$a_j(t+1) = \frac{a_j(t)}{R^s(t)} \tag{5}$$

$a_j(t)$ is the weight for the j-th index at the t-th search iteration. This procedure should be executed for all indices until the distance between the chosen and key sounds becomes the smallest.

2.4. Sound Display

It has been explained that the system yields five candidate sounds when a key sound is given as input. It is very important for the system to display the candidate sounds, so that the user can make a good selection. In the case of searching an image database, it will be easy for the user to select an image from a number of candidates when the database system presents them simultaneously. Unfortunately, unlike the visual representation, it will be impossible for the user to select one sound from a number of sound candidates that are simultaneously presented. In the proposed system, to help the user to make a choice, spatial and temporal effects are generated. Figure 7(a) shows the spatial effect generated by the system. In this figure, it is shown that 90% of Sound 1 is output to the left speaker, while 10% of the sound is directed toward the right speaker. For Sound 2, the composition becomes 70% and 30% and so on. This effect will give the impression to the user that different sounds are generated at different places. Figure 7(b) explains the temporal difference of each sound's generation, in which the system is set at a 0.2 s time delay between each sound presentation. The temporal and spatial effects will help the user to distinguish one sound from another and make a selection according to the user's preference.

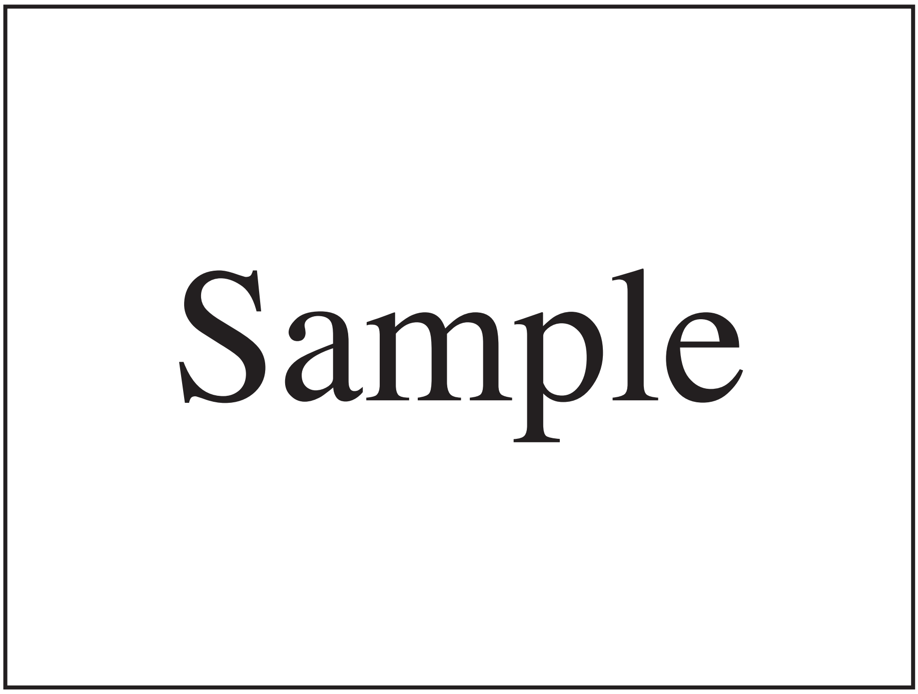

Fig. 7 Sound presentation.

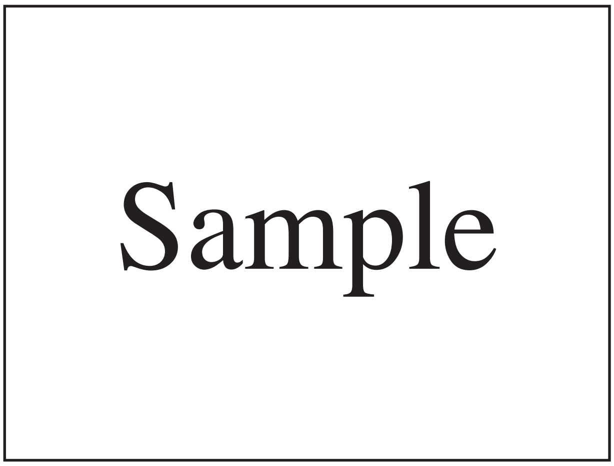

Fig. 8 GUI for the proposed system.

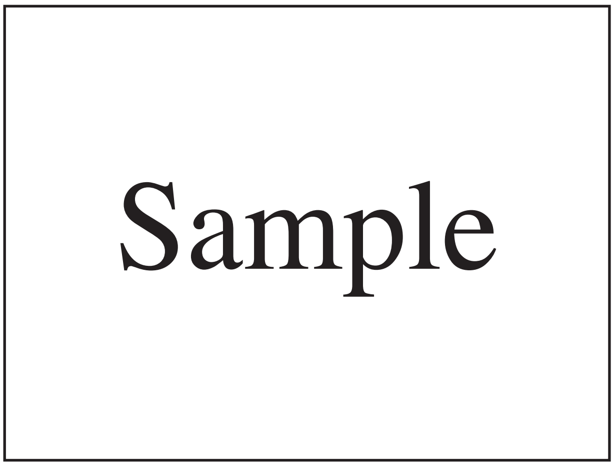

Fig. 9 Search iterations.

After the continuous presentation of each candidate, it is also possible for the users to hear the respective sound candidate before making a final decision. The GUI for sound presentation and key registration is shown in Fig. 8. The search query is first input with a microphone. The wave form of the query-key is shown in the first window in the left column in Fig. 8, with the rest of the windows showing the wave forms of 5 candidate sounds. Sounds and visual presentations will help the user to make a selection. The selected sound will be registered as a new key. This process will be iterated until the user finds the target sound.

The GUI for sound presentation and key registration is shown in Fig. 9. The search query is first input with a microphone. The wave form of the query-key is shown in the first window in the left column in Fig. 9, with the rest of the windows showing the wave forms of 5 candidate sounds. Sounds and visual presentations will help the user to make a selection. The selected sound will be registered as a new key. This process will be iterated until the user finds the target sound.

3. Experiments and results

The proposed system was implemented on a Windows 2000 machine (CPU: Pentium III: 600 kHz). It takes about 5 seconds to extract parameters from the key

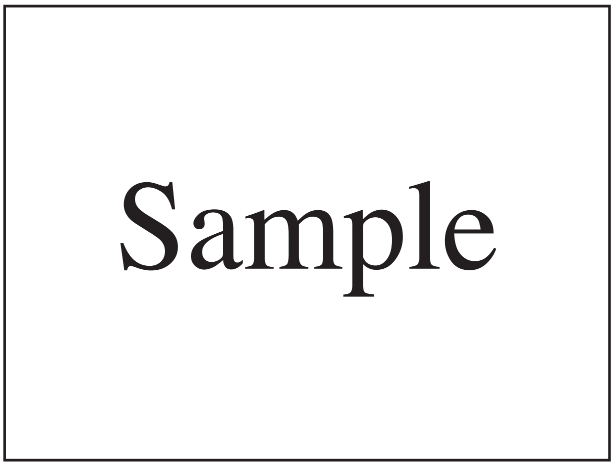

Fig. 10 Satisfaction index.

sound and less than 1 second to retrieve five of the most similar sounds from 1,800 sounds. To test the proposed system, experiments with 10 users with no specific musical expertise and technical knowledge about the proposed system were done. Each of the users was asked to input a sound and extract a sound that was similar to the input sound according to the user's preference. Each user conducted 5 searches using 5 different sounds as follows:

(1) Sound 1: whistle
(2) Sound 2: noise (exhaling into the microphone)
(3) Sound 3: human speech
(4) Sound 4: metallic bell
(5) Sound 5: crumpling a piece of paper

For comparison we also conducted experiments on the database system in which the indices for each sample in the database are 896 frequency components of each sound. In total, we conducted 4 types of experiments as follows,

(1) Type 1: 64-indexed data with adaptive preference weights.
(2) Type 2: 64-indexed data without preference weight.
(3) Type 3: 896-indexed data with adaptive preference weights.
(4) Type 4: 896-indexed data without preference weight.

To evaluate the quality of the proposed system, we asked the user to rank the quality of the extracted sound in 5 steps, in which 1 indicates "very poor" and 5 indicates "very good." The average of the rank is shown as "index" in Fig. 10. Figure 10 shows that, in general, the proposed system produces better searching results than other systems.

4. Conclusions and future work

A new type of sound database system using a sound as a key for data retrieval was proposed. For data indexing, we introduced 64 parameters that characterize the timbre of the sound. One of the advantages of the proposed database system is that the system does not require any prerequisite knowledge for the users. This advantage enables the users to search the database with their personal preferences. This characteristic will be very important for a large database system with a wide range of users, because it will be impossible for the database designer to encompass all the characteristics of the users in building the database system. The proposed system is intended as basic research for building a new database paradigm. Future research topics will include evaluation of the proposed system in a much larger database, the function for automatic new data inclusion, and the implementation of a sound modification function to create new sounds from the retrieved sound.

REFERENCES

[1] H. Qi, T. Muramatsu and S. Hashimoto, "Multimedia environment for sound database system," *Proc. ICMC 1997*, pp. 105–108 (1997).

[2] T. Blum, D. Keislar, J. Wheaton and E. Wold, "Audio database with content-based retrieval," *Annu. Rev. Physiol.*, **61**, 457–476 (1995).

[3] D. Keislar, T. Blum, J. Wheaton and E. Wold, "Audio analysis for content-based retrieval," *Proc. ICMC 1995*, pp. 199–202 (1995).

[4] R. Vertegaal and E. Bonis, "ISEE: An intuitive sound editing environment," *Comput. Music J.*, **18**(2), 21–29 (1994).

[5] B. Feiten and S. Gunzel, "Automatic indexing of a sound database using self-organizing neural nets," *Comput. Music J.*, **18**(3), 53–65 (1994).

[6] S. Middo, O. Kitamura, S. Namba and R. Matsumoto, "Research on timbre-II," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 55–56 (1961).

[7] S. Middo, O. Kitamura, S. Namba and R. Matsumoto, "Research on timbre-IV," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 65–66 (1962).

[8] K. Kashino, K. Nakadai, T. Kinoshita and H. Tanaka, "Application of Bayesian probability to music scene analysis," in *Computational Auditory Scene Analysis*, D. Rosenthal and H. Okuno, Eds. (Lawrence Erlbaum, Marwah, NJ, 1998), pp. 115–137.

[9] V. Openheim and S. Nawab, *Symbolic and Knowledge-Based Signal Processing* (Prentice Hall, Englewood Cliffs, NJ, 1992).

[10] T. Nakatani, H. Okuno and T. Kawabata, "Auditory stream segregation in auditory analysis with a multi-agent system," *Proc. 12th Natl. Conf. Artificial Intelligence*, pp. 100–107 (1994).

[11] D. Ellis, "A computer implementation of phsychoaccoustic grouping rules," *Proc. 12th Int. Conf. Pattern Recognition* (1994).

[12] S. Namba, S. Kuwano and T. Kato, "The relation between loudness and rise time as a function of energy," *J. Acoust. Soc. Jpn. (J)*, **30**, 144–150 (1974).

[13] J. R. Miller and E. C. Caterette, "Perceptual space for musical structures," *J. Acoust. Soc. Am.*, **58**, 711–720 (1975).

[14] K. Hirose, *Music Psychology* (Gakugei Shuppan-sha, Tokyo, 1983).

[15] S. Stevens, "On the psychological law," *Psychol. Rev.*, **64**, 153–181 (1957).

[16] R. Plomp, *Aspects of Tone Sensation—A Psychophysical Study* (Academic Press, London, 1976).

[17] A. de Bruijn, "Timbre-classification of complex tone," *Acustica*, **40**, 108–114 (1978).

[18] B. L. Pratt and P. E. Doak, "A subjective rating scale for timbre," *J. Sound Vib.*, **45**, 317–328 (1976).

[19] J. P. Guildford, *Psychometric Methods* (McGraw Hill, New York, 1954).

[20] C. E. Osgood, G. Suci and P. Tannenbaum, *The Measurement of Meaning* (University of Illinois Press, Urbana, 1957).

[21] S. Kondo and C. Hayashi, "On the preference of quality," *J. Acoust. Soc. Jpn. (J)*, **21**, 216–226 (1965).

[22] M. Saito and T. Tsumura, "Relation between timbre and depth of frequency fluctuation—Comparison between listening condition through headphone and free field—," *Tech. Rep. Mus. Acoust. Acoust. Soc. Jpn.*, MA90-5, pp. 3–10 (1990).

Hai Qi received a B.E. degree from Harbin Engineering University, China, in 1991, and an M.E. degree from Waseda University, Japan, in 1998. He is currently working toward his Ph.D. at Waseda University.

Pitoyo Hartono received his B.S and Ph.D. degrees from the Department of Applied Physics, Waseda University in 1993 and 2002, respectively. Since 2001, he has been a research associate with the Advanced Research Institute for Science and Engineering, Waseda University. His research interests are neural networks, evolutionary computation, and signal processing. He is a member of IEEE and INNS.

Kenji Suzuki received his B.S. and M.S. degrees from the Department of Applied Physics, Waseda University in 1997, 2000, respectively. Since 2000, he has been a Ph.D. candidate in Waseda University. He is currently a research associate in the Department of Applied Physics, Waseda University. His research interests include artificial neural networks, KANSEI information processing and robotics. He is a member of IEEE and ICMA.

Shuji Hashimoto received B.E, M.E., and Ph.D. degrees in Applied Physics from Waseda University in 1970, 1973 and 1977, respectively. Currently he is a Professor in the Department of

Applied Physics, Waseda University. His main research interests are image and sound processing. He is a member of ICMA, SICE, ISCIE, IPSJ, and the Robotics Society of Japan.

Cover Letter for Acoustical Science and Technology

(1) **Title of paper**
Sound database retrieved by sound
Sample of SUB-TITLE

(2) **Full name(s) of author(s)**
Hai Qi*, Pitoyo Hartono†, Kenji Suzuki‡ and Shuji Hashimoto§

(3) **Affiliation(s)**
Department of Applied Physics, Waseda University, Ohkubo 3–4–1, Shinjuku-ku, Tokyo, 169–8555 Japan

(4) **Approximately five keywords**
Sound database, Adaptive query, Intuitive data retrieval

(5) **PACS number**
43.60.Lq, 43.66.Jh

(6) **Short running title**
H. QI *et al.*: SOUND DATABASE RETRIEVED BY SOUND

(7) **Category of article:**
PAPER

(8) **Mailing address**
東京都文京区本郷7-3-1
東京大学
東大太郎

(9) **Classification**
Speech

(10) **Number of pages**
TEXT: 9 FIGURES: 10 TABLES: 1