Relationships between auditory impressions and onomatopoeic features for environmental sounds

Masayuki Takada\textsuperscript{1,*}, Kazuhiko Tanaka\textsuperscript{2} and Shin-ichiro Iwamiya\textsuperscript{1}

\textsuperscript{1}Department of Acoustic Design, Faculty of Design, Kyushu University, 4–9–1 Shiobaru, Minami-ku, Fukuoka, 815–8540 Japan
\textsuperscript{2}Honda R&D Co., Ltd., Asaka R&D Center, 3–15–1 Senzui, Asaka, 351–8555 Japan

(Received 16 December 2004, Accepted for publication 5 October 2005)

Abstract: In order to clarify the features of environmental sounds, psychoacoustical experiments using onomatopoeic representations were conducted. The onomatopoeic representations obtained for each sound and participant were expressed using phonetic parameters, such as place, manner of articulation, and vowels. Onomatopoeic representations were classified based on similarities of phonetic parameters using a hierarchical cluster analysis. As a result, similar acoustic properties were identified in the stimuli expressed by onomatopoeic representations classified into the same clusters. Furthermore, the auditory impressions associated with stimuli were measured by a semantic differential method using 13 adjective pairs. Factor analysis was applied to the average ratings for each sound for each scale. The three factors obtained were the emotion factor, the clearness factor, and the powerfulness factor. From these results, the relationships among the acoustic properties of sound stimuli, the impressions associated with them, and their onomatopoeic features were discussed. For example, onomatopoeic representations including voiced consonants were frequently used for sounds displaying components in the frequency region below about 1 kHz in their spectrum, inducing ‘‘powerfulness’’ and ‘‘darkness, dullness, and mudliness’’ impressions. Furthermore, similar relationships were found in supplementary experiments using various band noises.

Keywords: Onomatopoeic representation, Environmental sound, Factor analysis, Phonetic feature, Acoustic property, Auditory Impression

PACS number: 43.66.Jh [DOI: 10.1250/ast.27.67]

1. INTRODUCTION

Sounds infinite in variety surround us throughout our lives. When we report on sounds to others in our daily lives, onomatopoeic representations are frequently used. Onomatopoeic representations may be related to the acoustic properties of sounds. On the other hand, the acoustic properties of sounds may also induce auditory impressions in a listener. Thus there may be relationships between onomatopoeic representations and the auditory impressions associated with sounds. In a number of previous studies, the relationships between the acoustic properties of sounds (in the time and frequency domain) and onomatopoeic features have been investigated [1–3]. However, only a few studies have touched on the relationship between the auditory impressions of sounds and onomatopoeic representations. Tanaka et al. provided verbal descriptions (including onomatopoeic representations) for sounds and vibrations inducing comfortable sensations [4]. Their verbal descriptions, however, were not derived from experiments. Thus, the relationship between auditory impressions induced by the acoustic properties of sounds and the associated onomatopoeic representations remains insufficiently discussed. Such research is important since it would allow one to measure the impressions of sounds without conducting standard psychoacoustic experiments such as those using the semantic differential method.

Our two previous studies confirmed the validity of using onomatopoeic representations to identify the acoustic properties of sounds of manual operation emitted from office equipment and audio signals emitted from domestic electronic products [5,6]. Furthermore, the relationships between the onomatopoeic features of sounds of manual operation, audio signals, and subjective impressions, such as product imagery and functional imagery, were discussed. In those experiments, however, stimuli were
limited to sounds emitted from machinery.

In the present study, psychoacoustic experiments were conducted to investigate the possibility of using onomatopoeic representations to estimate the impressions associated with various environmental sounds. Initially, a free description experiment using onomatopoeic representations was carried out. Subsequently, subjective evaluation experiments regarding familiar auditory impressions, such as powerfulness and sharpness associated with the same sound stimulus, were examined. The relationships among acoustic properties, impressions associated with sounds, and onomatopoeic features were then discussed. Furthermore, in order to verify the obtained knowledge, supplementary experiments were conducted using synthesized sounds.

### 2. EXPERIMENT 1: EXPERIMENTS WITH ENVIRONMENTAL SOUNDS

#### 2.1. Experimental Methods

Thirty-six sounds selected from commercially available compact discs [7] were used as the sound stimuli. These comprised various natural environmental sounds as well as synthesized sounds such as harmonic complex tones, a siren, and white noise, as shown in Table 1. In Table 1, the supplementary explanations for some stimuli are described within parentheses. These sounds were selected from the viewpoint that we often hear them both outdoors and indoors in our daily lives. Although they do not cover all ambient sounds, they include many sounds familiar to Japanese, such as machinery noise (e.g., cameras and jackhammers), traffic noise (e.g., trains and cars), natural

<table>
<thead>
<tr>
<th>No.</th>
<th>Name of stimuli</th>
<th>Examples of onomatopoeic representations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>whizzing sound (similar to the crack of a whip)</td>
<td>/fAN/  [fæ:v], /huyAN/  [cʰiːm]</td>
</tr>
<tr>
<td>2</td>
<td>idling sound of a diesel engine</td>
<td>/gagagaga在这个词应该读作“gaga”/ [gaːɡa, ɡaɡa], /burogorororo/ [bʊroɡoɾoɾo]</td>
</tr>
<tr>
<td>3</td>
<td>sound of water dripping</td>
<td>/pokaQ/  [pɔ:ʔ], /potaN/  [pɔtaʔ]</td>
</tr>
<tr>
<td>4</td>
<td>bark of a dog (barking once)</td>
<td>/waN/  [waʔ], /wa/  [waʔ]</td>
</tr>
<tr>
<td>5</td>
<td>ring of a telephone</td>
<td>/piriririri/  [piɾiɾiɾiɾiɾi], /pirororororo/ [piɾoɾoɾoɾoɾo]</td>
</tr>
<tr>
<td>6</td>
<td>owl hooting</td>
<td>/kururu/  [kɯɾuɾuɾu], /fururuuru/  [fuɾuɾuɾuɾu]</td>
</tr>
</tbody>
</table>
| 7   | vehicle starter sound | /gyuruuruuruN/  [ɡjɯɾuɾuɾuɾuɾu], /kururuuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruuruureru
sounds (e.g., water, wind, and thunder), sounds generated by living things (e.g., animals such as a dog and a sparrow, and insects), sounds generated by phenomena (e.g., sounds caused by bumping and friction between objects), and synthesized sounds used as auditory signals. Gaver stated that sound-producing events might be divided into three general categories: vibrating solids, gasses, and liquids, in his discussion on the ecological approach to auditory perception [8]. The sound stimuli in the present study comprised not only sounds that could be classified into these three categories, but also synthesized sounds and sounds generated by living things. Moreover, in previous studies concerning environmental sounds, similar sounds were also used as stimuli [9–15]. All stimuli were transferred from audio compact discs into the hard disk of a computer (IBM NetVista A30p). The order of presentation of the stimuli was randomized for each subject. The single event sound exposure level \( L_{AE} \) of each stimulus was measured using the ear simulator (Brüel&Kjær Type 4153) and the sound level meter (RION NA-29) because some of the sound stimuli had durations shorter than 1,000 ms. The mean \( L_{AE} \) of these stimuli was 51 dBA in the headphones. In these experiments the participants were able to listen to the sound stimuli for as many times as they felt necessary.

### 2.2. Results

#### 2.2.1. Analysis of onomatopoeic representations

In Table 1, examples of the onomatopoeic representations obtained are shown for each stimulus used in the experiment. To examine the onomatopoeic features necessary to represent the acoustic properties of sound stimuli, the onomatopoeic representations were coded using 24 phonetic parameters consisting of 7 places of articulation (labio-dental, bilabial, alveolar, post-alveolar, palatal, velar, and glottal), 6 manners of articulation (plosive, fricative, nasal, affricate, approximant, and flap) [16], 5 Japanese vowels (/a/, /i/, /u/, /e/, /o/), voiced and voiceless consonants, syllabic nasals, geminate obstruents, palatalized consonants, and long vowels. Furthermore, onomatopoeic representations were classified based on the similarities of 20 phonetic parameters (see Table 3) using a hierarchical cluster analysis (the ward method). In this analysis, the measure of the similarities was the Euclidean distance. The four parameters of labio-dental, glottal, nasal, and vowel /e/ were eliminated in the cluster analysis since these parameters rarely appeared in the onomatopoeic representations. Figure 1 shows the results of the cluster analysis. In this figure, “Rescaled Distance Cluster Combine” means that the distances among the clusters were rescaled to values between 0 and 25 [17]. We interpreted the onomatopoeic representations as being roughly classified into 5 clusters. In each cluster, the medians of the number of moras across all participants’ onomatopoeic representations for all stimuli are also indicated.

#### 2.2.2. Analysis of subjective rating

The rating values were averaged for each scale and for each sound, and factor analysis was applied to the averaged values. As a result of taking into account the factors for which the eigenvalues were more than 1, a three-factor solution was obtained. The eigenvalues were 6.29, 3.01, and 1.40. The first, second, and third factors account for 48.4%, 23.1%, and 10.8% of the total variance in the data, respectively. Finally, the factor loadings of each factor on each sound stimulus had durations shorter than 1,000 ms. The mean \( L_{AE} \) of these stimuli was 51 dBA in the headphones. In these experiments the participants were able to listen to the sound stimuli for as many times as they felt necessary.

<table>
<thead>
<tr>
<th>Pair of adjectives</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>tasteful–tasteless</td>
<td>0.878</td>
<td>0.057</td>
<td>0.203</td>
</tr>
<tr>
<td>desirous of hearing–</td>
<td>0.833</td>
<td>0.293</td>
<td>0.332</td>
</tr>
<tr>
<td>undesirous of hearing–</td>
<td>0.823</td>
<td>−0.174</td>
<td>0.185</td>
</tr>
<tr>
<td>rural–urban</td>
<td>0.710</td>
<td>0.505</td>
<td>0.360</td>
</tr>
<tr>
<td>pleasant–unpleasant</td>
<td>0.635</td>
<td>−0.068</td>
<td>0.491</td>
</tr>
<tr>
<td>slow–fast</td>
<td>0.585</td>
<td>0.153</td>
<td>−0.009</td>
</tr>
<tr>
<td>soft–hard</td>
<td>−0.200</td>
<td>−0.850</td>
<td>−0.382</td>
</tr>
<tr>
<td>bright–dark</td>
<td>−0.025</td>
<td>0.834</td>
<td>−0.048</td>
</tr>
<tr>
<td>smooth–rough</td>
<td>0.225</td>
<td>0.735</td>
<td>0.301</td>
</tr>
<tr>
<td>sharp–dull</td>
<td>−0.519</td>
<td>0.575</td>
<td>−0.290</td>
</tr>
<tr>
<td>modest–loud</td>
<td>0.385</td>
<td>−0.179</td>
<td>0.874</td>
</tr>
<tr>
<td>strong–weak</td>
<td>−0.285</td>
<td>−0.391</td>
<td>−0.823</td>
</tr>
<tr>
<td>powerful–powerless</td>
<td>−0.060</td>
<td>−0.473</td>
<td>−0.808</td>
</tr>
</tbody>
</table>
factor because adjective pairs such as “muddy/clear” and “bright/dark” have high factor loadings. The third factor is interpreted as the powerfulness factor because the adjective pairs “modest/loud,” “strong/weak,” and “powerful/powerless” have high factor loadings.

In order to clarify the features of the sound stimuli with regard to the factors, factor scores for each stimulus for each factor were also obtained.

### 2.3. Discussion

#### 2.3.1. Relationships between the onomatopoeic features and acoustic properties of environmental sounds

In order to discuss the relationship between the onomatopoeic features and the acoustic properties of sound stimuli, the medians of the number of phonetic parameters in the onomatopoeic representations described by all of the participants were calculated in each cluster, as shown in Table 3. For each cluster, the prominent relationships between the acoustic properties of sound stimuli and the onomatopoeic features are discussed below.

Cluster 1 was comprised of 10 sound stimuli (numbers 8, 32, 13, 1, 4, 14, 17, 27, 34, and 18, i.e., “hand clap,” “bark of a dog,” “sound of a wind chime,” etc.). The mean duration of all stimuli in cluster 1 (329 ms) was shorter than those in other clusters. These short stimuli can be represented onomatopoeically with only a few moras. Therefore, in cluster 1, the median of the number of moras in cluster 1 is the smallest of all five of the clusters, as shown in Fig. 1. The rank correlation coefficient (Spearman’s rho) between the median of the number of moras across all participants’ onomatopoeic representations and the duration of stimuli was statistically significant at \( p < 0.01 \) (\( r_s = 0.775, d.f. = 34, t = 7.15 \)). Examples of the obtained onomatopoeic representations are /paN/ [pa] (for “hand clap”), /waN/ [wan] (for “bark of a dog”) and /riN/ [rin] (for “sound of a wind chime”).

Two subdivisions can be found in cluster 1, as shown in Fig. 1. The five stimuli of “sweeping tone (No. 14),” “twittering of a sparrow (No. 17),” “bell of a microwave (No. 27),” “sound of a wind chime (No. 34),” and “harmonic complex tone (No. 18),” which are classified in one subdivision of cluster 1, displayed high energy in the high-frequency region of the spectrum. Spectral centroids of these stimuli are above 2 kHz (No. 14: 2,125 Hz, No. 17: 4,423 Hz, No. 18: 4,229 Hz, No. 27: 5,476 Hz, No. 34: 7,121 Hz). All participants used vowel /i/ in their onomatopoeic representations for these five stimuli. The second highest median of the number of vowel /i/ in cluster 1 (5.5, see Table 3) is due to these five stimuli being classified in the subdivision of cluster 1.

In order to examine the relationship of the spectral feature between Japanese vowels and the sounds represented using them, three male speakers produced the five vowels three times, and then the mean spectral centroids of vowels were compared for each speaker, as shown in Table 4. Consequently, for each speaker, the spectral centroids of vowels /i/ and /o/ were the highest and lowest of all the five vowels, respectively. Therefore, vowel /i/ might be used in onomatopoeic representations for sounds displaying high-frequency components. Similar results were also obtained in previous studies [1,2]. Figure 2 shows the spectrogram for the stimulus of “twittering of a sparrow” as a typical stimulus in cluster 1. This figure shows the SPL contour in the frequency and time domains. The dark area represents high energy. The stimulus of “twittering of a sparrow” displayed high energy in the high-frequency region above 2 kHz. A typical onomatopoeic representation for this sound is /pijo/ [pijo].

Cluster 2 was comprised of nine sound stimuli (numbers 10, 25, 30, 35, 9, 20, 22, 24, and 31, i.e., “sound of a waterfall,” “vehicle horn,” “sound of a trumpet,” etc.). The mean duration of the stimuli in cluster 2 was the longest of all the clusters (2,524 ms). However, in this cluster, the median of the number of moras was not the highest of all clusters, as shown in Fig. 1. These stimuli displayed temporally monotonous characteristics of the spectral distribution for the greater part of their duration. Table 3 shows that long vowels are frequently used in onomatopoeic representations to represent this acoustic
property. Examples of the obtained onomatopoeic representations are /zaaaaa/ [dzär] (for “sound of a waterfall”) and /puuuuuuN/ [puuz] (for “sound of a trumpet”).

Furthermore, Table 3 shows that vowels /a/ and /u/ were used more frequently than vowel /i/ (or /o/) in cluster 2. Figure 3 shows the spectrogram for the stimulus of “vehicle horn” as a stimulus represented using these onomatopoeic features. A typical onomatopoeic representation for this sound is /puu/ [puuz].
Cluster 3 was comprised of the two sound stimuli of “ring of a telephone (No. 5)” and “chirping of an insect (No. 16).” The mean duration of these two stimuli in cluster 3 was 1,206 ms. These stimuli displayed fluctuations in the temporal envelope. The modulation rates of these stimuli were about 11 Hz and 45 Hz, respectively. According to Table 3, the numbers of flap sounds and voiceless consonants in cluster 3 were the largest in all of the clusters. Moreover, the median of the number of moras in cluster 3 was the second highest of the clusters, as shown in Fig. 1. Fast fluctuations in the temporal envelope tend to be represented onomatopoeically with repeated flap consonants such as /piriririri/ [piɾiɾiɾiɾiɾi] (for “ring of a telephone”) and /kyuriririririi/ [kʰyɾiɾiɾiɾiɾi] (for “chirping of an insect”). Iwamiya and Nakagawa also reported that amplitude-modulated audio signals with modulation rates of about 12 to 37 Hz were represented using flap consonants [2].

Additionally, the sound stimulus “chirping of an insect” in cluster 3 displayed components in the high-frequency region above 4 kHz in its spectrum. Figure 4 shows the spectrogram for this stimulus. To represent the acoustic properties of fluctuations in the temporal envelope and high-frequency components displayed by this stimulus, the vowel /i/ is also frequently used together with a repeated flap consonant. The largest median of vowel /i/ in cluster 3 was due to these properties (see Table 3).

Cluster 4 was comprised of 11 sound stimuli (numbers 6, 7, 21, 33, 29, 36, 3, 19, 15, 11, and 23, i.e., “owl hooting,” “typing sound,” “footsteps,” etc.). The median of the number of moras in cluster 4 was the second smallest of all the clusters, as shown in Fig. 1. This may be due to the characteristic that the mean duration of the sound stimuli in cluster 4 was the second shortest of all the clusters (722 ms). Furthermore, these stimuli were comprised of intermittent sounds. Figure 5 shows the spectrogram for the stimulus “typing sound” as a typical stimulus of cluster 4. The short components of intermittent sounds of the stimuli in cluster 4 are represented onomatopoeically with plosives of voiceless consonants. Table 3 shows that a relatively large number of plosives and voiceless consonants are used in onomatopoeic representations for stimuli classified in cluster 4. Examples of the obtained onomatopoeic representations are /kurururu/ [kuɾuɾuɾuɾu] (for “owl hooting”), /katakoto/ [katakoto] (for “typing sound”), and /katsukotsu/ [katsukotsu] (for “footsteps”). From the viewpoint of phonetics, a plosive is produced when the exhalation is abruptly released from the oral cavity with high pressure. Therefore, a plosive can be suitable for representing the abrupt beginning of components of intermittent sound.

Cluster 5 was comprised of the four sound stimuli “thunderclap (No. 26),” “sound of a passing train (No. 28),” “sound of a noisy construction site (No. 12),” and “idling sound of a diesel engine (No. 2).” These stimuli were comprised of intermittent sounds similar to the stimuli of cluster 4. The mean duration of the sound stimuli in cluster 5 was the second longest of all the clusters (1,966 ms). However, in order to represent the many short components of intermittent sounds, onomatopoeic representations with the largest number of moras were used, as shown in Fig. 1. Table 3 also indicates that these stimuli are represented using many plosives to represent intermittent sounds. In particular, plosives of voiced consonants are frequently used. Examples of the onomatopoeic representations obtained are /gakanaɡakaN-gakanaɡakaN/ [ɡakaN ɡakajɑjɑkan] (for “sound of a passing train”) and /gagagagagagagagaga/ [ɡaŋaŋaŋaŋaŋa-
shown in Table 3. The stimuli contained intermittent sounds displaying energy in the frequency range from about 50 Hz to 8 kHz. Figure 6 shows the spectrogram for the stimulus “sound of a noisy construction site” as a typical stimulus of cluster 5. In order to measure the frequency region of the dominant components of the sound stimulus in its spectrum, the lowest and highest cut-off frequencies in the frequency band 10 dB below the spectral peak were computed. The mean frequencies of the lowest and highest frequency components in the aforementioned frequency band of sound stimuli in cluster 5 were 65 Hz and 1,077 Hz, respectively. Accordingly, sound stimuli in cluster 5 displayed dominant components in the low-frequency region. In onomatopoeic representations for all the stimuli in cluster 5, /ga/ was frequently included. Three male speakers produced /ga/ three times, and then, in the same manner as above, the lowest and highest cut-off frequencies in the frequency band 10 dB below the spectral peak were computed for the spectrum of voiced consonant /g/. Table 5 shows the mean frequencies of the lowest and highest components in the aforementioned frequency band for each speaker. Although the average lowest and highest cut-off frequencies varied widely among speakers, the dominant frequency band of this voiced consonant was in the frequency region below 1.2 kHz. Therefore, a voiced consonant may be frequently used to represent the sound stimuli in cluster 5, which mainly displayed energy in the low-frequency region, as shown in Table 3.

As a result of a cluster analysis for onomatopoeic representations based on the similarities of phonetic parameters, similar acoustic properties were found in sound stimuli grouped together in the dendrogram (the result of the cluster analysis). The relationships between the acoustic properties and the onomatopoeic features for each cluster were discussed. These results indicate that onomatopoeic representations can be effective for identifying the various acoustic properties of environmental sounds.

2.3.2. Relationships between the onomatopoeic features and auditory impressions associated with environmental sounds

The results of a cluster analysis confirmed the validity of using onomatopoeic representations to identify the acoustic properties of environmental sounds. Since these acoustic properties induce auditory impressions, it follows that there should be onomatopoeic representations that can serve to express the auditory impressions associated with particular sounds. In order to clarify the relationships between onomatopoeic features and auditory impressions, the rank correlation coefficient (Spearman’s rho) between the numbers of each phonetic parameter in onomatopoeic representations described by all of the participants and the factor scores for each factor was computed. T-tests were used to confirm the significance of the rank correlation coefficients.

In the adjective scales with high factor loadings on the emotion factor, aspects other than the acoustic properties of sounds, such as the meanings of sounds, might affect auditory impressions. The sound stimuli with highly negative factor scores of the emotion factor (inducing the impressions of “tasteful, rural, and pleasant”) included “sound of a temple bell,” “sound at a beach,” “sound of a flowing stream,” and “sound of a wind chime.” These sounds are typical of sounds generally preferred by the Japanese (e.g., [12]). The “emotional” impression of these sounds is thought to be affected by the sense of the seasons or customs peculiar to Japan rather than the acoustic properties of the sounds themselves. Furthermore, the consistency of the subjective evaluations of these sounds with regard to how well they are known and how much they are preferred among the Japanese suggests that the participants could recognize the sources of these sounds. Consequently, no clear relationship between onomatopoeic features and auditory impressions was found regarding the emotion factor. Therefore, in the following analysis, we discuss the clarity and powerfulness factors.

Table 6 presents the phonetic parameters that show that the rank correlation coefficients are statistically significant at $p < 0.01$ regarding the clarity and powerfulness factors. Figure 7 shows the factor scores of sound stimuli for the clarity factor (the second factor) and powerfulness factor (the third factor).

The rank correlation coefficients between the clarity factor and the numbers of velar, alveolar, vowel /o/, and voiced consonant are statistically significant, as shown in

![Fig. 6 Spectrogram of stimulus “sound of a noisy construction site” classified in cluster 5.](image)
Therefore, vowel /o/ is frequently used to represent sounds displaying energy in the high-frequency region and dullness in subjective evaluations. Conversely, vowel /i/ (0.429, d.f. = 34, t = 2.77) and voiced consonant (0.554, d.f. = 34, t = 3.88) were used in onomatopoeic representations for stimuli with the number of vowel /i/ indicating that vowel /i/ is frequently used to represent sound stimuli inducing impressions of “brightness and sharpness.”

Almost all of the participants used vowel /i/ in their onomatopoeic representations of seven stimuli (numbers 14, 16, 17, 18, 23, 27, and 34). On the other hand, in onomatopoeic representations of five stimuli (numbers 3, 15, 19, 21, and 25), vowel /o/ was used by almost all of the participants. According to Table 6, the positive correlation of the clearness factor with the number of vowel /i/ was statistically significant at p < 0.05 (r = -0.398, d.f. = 34, t = 2.53). Vowel /i/, with the highest spectral centroid among the Japanese vowels, is frequently used to represent sound stimuli inducing impressions of “brightness and sharpness.”

The positive correlation of the clearness factor with the number of voiced consonant also indicates that voiced consonants are frequently used in onomatopoeic representations of stimuli inducing impressions of “dullness,” “darkness,” and “muddiness.”

Regarding voiced consonant, the rank correlation coefficient between the powerfulness factor and the number of voiced consonant was also statistically significant, as shown in Table 6. This result indicates that voiced consonants are used to represent sounds inducing impressions of “powerfulness” as well as impressions of “dullness,” “darkness,” and “muddiness.”

Almost all of the participants used voiced consonants in onomatopoeic representations of 10 stimuli (numbers 2, 7, 21, 25, 26, 28, 33, and 35). According to the results of a cluster analysis, these stimuli are classified into cluster 5 and the subdivisions of clusters 2 and 4. Figure 7 shows that the above 10 stimuli represented using voiced consonants by all of the participants induce not only impressions of “dullness,” “darkness,” and “muddiness,” but also impressions of “powerfulness.” Conversely, no one used voiced consonants to represent 19 stimuli (numbers 1, 3, 4, 5, 6, 8, 13, 14, 15, 17, 18, 19, 20, 22, 27, 29, 31, 32, and 34). To measure the frequency region of the dominant
components of the sound stimulus in its spectrum, the lowest and highest cut-off frequencies in the frequency band 10 dB below the spectral peak were computed. In the former group (represented using a voiced consonant), the mean frequencies of the lowest and highest frequency components in the aforementioned frequency band are 140 Hz and 1,100 Hz, respectively. In the latter group (not represented using a voiced consonant), these frequencies are 1,213 Hz and 2,892 Hz, respectively. The dominant frequency band of sound stimulus represented using a voiced consonant is lower than that not represented using a voiced consonant.

In phonetics, the relationships between the plosives and the spectral shapes were previously identified [19]. For example, the plosive consonant articulated at alveolar (/d/ or /t/) displays the flat spectral shape or energy in the high-frequency region above 4 kHz in its spectrum, and the plosive consonant articulated at velar (/g/ or /k/) displays energy in the intermediate frequency region from 1.5 kHz to 4 kHz. However, regarding these relationships, the difference in the spectral shape between the voiced and voiceless consonants (e.g., /g/ and /k/) has not been discussed. By grouping the stimuli into those represented with or without voiced consonants, we see that the voiced and voiceless consonants may display their own spectral features. Table 5 shows the lowest and highest frequency components in the aforementioned frequency band for the voiceless consonant /k/ as well as the voiced consonant /g/. /k/ was extracted from the speech sound /ka/ produced by three male speakers. According to Table 5, the dominant frequency band in the spectrum of voiced consonant /g/ is lower than that of voiceless consonant /k/. These results suggest that a voiced consonant can be used for the sound displaying energy mainly in the frequency region below about 1 kHz, and a voiceless consonant can be used for the sound displaying the components mainly in the frequency region above 1 kHz. As mentioned in the previous section, the sound stimuli in cluster 5, which were onomatopoeically represented using voiced consonants, displayed the dominant components in the frequency region below about 1 kHz. Iwamiya and Nakagawa also reported that the fundamental frequency and the frequencies of dominant components of the audio signals represented using voiceless consonants were higher than those represented using voiced consonants [2].

In subjective evaluations, the acoustic property of components in the low-frequency region generally might induce impressions of “dullness” and “dullness” [18]. The actual positive rank correlation coefficient between the clearness factor and the number of voiced consonant was statistically significant, as shown in Table 6. Furthermore, 10 stimuli, represented using voiced consonants by all of the participants, were evaluated as being strong in terms of their impressions of “muddiness” (mean score is 2.85 on the adjective scale of “muddy/clear”).

According to Table 6, voiced consonants are also used to represent sound stimuli that induce an impression of “powerfulness.” These sound stimuli, which were onomatopoeically represented using voiced consonants, displayed the dominant components in the frequency region below about 1 kHz. Solomon’s studies concerning the timbre of sonar sounds reported that a factor interpreted as the “magnitude” dimension significantly correlated with the sound pressure level in the octave bands of 150–300 Hz, 300–600 Hz, and 600–1,200 Hz [20,21]. Therefore, the obtained relationship between voiced consonants, used for sound stimuli displaying energy in the frequency region below about 1 kHz, and the “powerfulness” impression may be possible. On the other hand, the stimuli in cluster 5, for which the participants used voiced consonants, might be generally heard at high volume in real situations. There is also some possibility that the imagery for the sound source would affect the correlation between the “powerfulness” impression and the number of voiced consonant.

In the onomatopoeic representations of sounds inducing impressions of “dullness,” “darkness,” and “muddiness,” velar and alveolar, which are places of articulation, are also frequently used. In the representations of short components in intermittent sounds and temporally fluctuating sounds, velar and alveolar are frequently used together with plosive and flap sounds, which are both manners of articulation. By using a combination of velars and plosives, the consonants /k/ and /g/, used in onomatopoeic representations such as /katakoto/ [katakoto], are articulated. Table 3 indicates that velars are frequently used in cluster 5, in which sound stimuli displaying intermittent sounds are classified. Using the combination of alveolar and flap sounds, consonant /t/ is used to represent temporally fluctuating sounds such as /burorororo/ [buɾuɾoɾoɾoɾo] (for “idling sound of a diesel engine”). Table 3 shows that alveolar sounds are frequently used in cluster 3, in which sound stimuli displaying temporally fluctuating sounds are classified. From these results, relationships between the onomatopoeic features of vowels (/i/ and /o/), voiced consonants, places of articulation (velar and alveolar), and auditory impressions of both the clearness and powerfulness factors associated with environmental sounds were suggested.

3. EXPERIMENT 2: SUPPLEMENTARY EXPERIMENTS WITH SYNTHESIZED SOUNDS

In Experiment 1, the relationships among acoustic properties, onomatopoeic features using Japanese vowels and voiced consonants, and auditory impressions of the clearness and powerfulness factors were discussed. In
Experiment 2, therefore, in order to verify these relationships, sound stimuli displaying acoustic properties similar to those represented onomatopoeically with Japanese vowels and voiced consonants for environmental sounds were synthesized. Psychoacoustic experiments were then carried out using these synthesized sounds.

### 3.1. Experimental Methods

In Experiment 1, the relationships among acoustic properties in the frequency domain, onomatopoeic features, and auditory impressions were mainly investigated. In order to verify these relationships, low-pass noise (LP), high-pass noise (HP), and band-pass noise (BP) were used as the sound stimuli in this experiment. The cut-off frequencies of low-pass noise, high-pass noise, and band-pass noise are shown in Table 7. The attack time of the stimulus is also shown in this table. The decay time is kept constant at 30 ms through all stimuli, and these stimuli have no steady-state parts. Therefore, the duration of a stimulus is equal to the sum of the attack time and the decay time. All stimuli were generated digitally using a computer (16-bit resolution, 44.1 kHz sampling rate).

Two types of experiments were carried out in the same manner as in Experiment 1: a free description experiment using onomatopoeic representations and a measurement of impressions associated with sound stimuli using adjective scales. Experiment 1 concerned the relationships among the acoustic properties, onomatopoeic features, and auditory impressions related to clearness and powerfulness factors. Therefore, the adjective scales of “muddy/clear,” “bright/dark,” “smooth/rough,” and “sharp/dull,” all of which had high loadings for the clearness factor, and “modest/loud,” “strong/weak,” and “powerful/powerless,” all of which had high loadings for the powerfulness factor, from Experiment 1 were used. The experimental equipment was the same as that used in Experiment 1. The 16 stimuli shown in Table 7 were presented in random order to each subject. The single event sound exposure level ($L_{AE}$) of these stimuli was approximately 60 dBA in the headphones (STAX Lambda Nova), and the standard deviation was less than 0.5 dBA. The measurement of $L_{AE}$ for each stimulus was carried out using the same apparatus as used in Experiment 1. Twelve participants (5 males and 7 females) with normal hearing ability participated.

### 3.2. Results

The obtained onomatopoeic representations were coded using 24 phonetic parameters in the same manner as in Experiment 1. The purpose of this experiment was to verify the relationships between the acoustic properties and onomatopoeic features of the Japanese vowels and voiced/voiceless consonants that were obtained in Experiment 1. Therefore, the numbers of these phonetic parameters in all of the participants’ onomatopoeic representations were compared among sound stimuli, as shown in Table 7.

The subjective rating values were averaged for each scale and for each sound stimulus.

In the following discussion, as mentioned above, the focus will be on the relationships among the eight frequency characteristics related to low-pass noise, high-pass noise, and band-pass noise, onomatopoeic features, and auditory impressions.
3.3. Discussion

3.3.1. Relationships between the onomatopoeic features and acoustic properties of synthesized sounds

In order to discuss the relationships between the onomatopoeic features and acoustic properties of sound stimuli, the numbers of Japanese vowels and voiced/voiceless consonants for each stimulus are shown in Table 7.

According to Table 7, voiced consonants were frequently used for all low-pass noises. In Experiment 1, sound stimuli represented using voiced consonants mainly displayed energy in the frequency region below about 1 kHz. In Experiment 2, voiced consonants were also used for low-pass noises with the higher cut-off frequencies of 5 and 10 kHz. On the other hand, the number of voiceless consonants slightly increased as the cut-off frequency of low-pass noise became higher than 1 kHz. Furthermore, voiceless consonants were frequently used for two types of high-pass noises displaying components in the high-frequency range above 5 kHz (HP2 and HP4). For high-pass noises with a cut-off frequency of 1 kHz (HP1 and HP3), voiceless consonants were shown to be used just as often as voiced consonants. Regarding the onomatopoeic representations for stimuli displaying the components in the frequency region higher than about 1 kHz, the number of voiceless consonants also tended to increase. For all band-pass noises, voiceless consonants were frequently used. In particular, although stimuli BP1 and BP4 displayed components in the low-frequency region below 1 kHz, both voiceless consonants and voiced consonants were used for these stimuli. Considering all of these tendencies, it seems that a voiced consonant may be more frequently used for sound displaying energy in the whole frequency region below about 1 kHz.

Vowel /i/ was frequently used to represent high-pass and band-pass noises displaying components in the frequency region higher than 4 kHz, such as HP2, HP4, BP3, and BP6, as shown in Table 7. On the other hand, in Experiment 1, the mean spectral centroid in the group of environmental sounds represented using vowel /i/ was 5,308 Hz. Furthermore, vowel /o/ was frequently used to represent low-pass and band-pass noises displaying components in the frequency region lower than 1 kHz, such as LP1, LP4, BP1, and BP4, as shown in Table 7. In Experiment 1, the mean spectral centroid in the group of environmental sounds represented using vowel /o/ was 1,569 Hz. Although this spectral centroid is slightly higher than the frequency components of synthesized sounds represented using vowel /o/ in Experiment 2, the relationships between the frequency contents of the sounds and vowels /i/ and /o/ in Experiment 2 are nearly consistent with those obtained from Experiment 1.

3.3.2. Relationships between the onomatopoeic features and auditory impressions associated with synthesized sounds

In order to clarify the relationships between onomatopoeic features and auditory impressions associated with synthesized sounds, the rank correlation coefficient (Spearman’s rho) between the numbers of each phonetic parameter (vowel /i/, vowel /o/, and voiced/voiceless consonants) on onomatopoeic representations described by all of the participants and the mean scores for each adjective scale was computed. T-tests were used to examine the significance of the rank correlation coefficients. The adjective scales that correlated significantly (at \( p < 0.01 \)) with vowels /i/ and /o/ and voiced/voiceless consonants are shown in Table 8. Vowels /i/ and /o/ correlated significantly with the adjective scales of “bright/dark” and “sharp/dull.” These correlation coefficients indicate that vowel /i/ is frequently used to represent stimuli inducing impressions of “brightness” and “sharpness,” and that vowel /o/ is frequently used to represent stimuli inducing impressions of “darkness” and “dullness.” These relationships are consistent with those obtained in Experiment 1.

On the other hand, voiced and voiceless consonants correlated significantly with the adjective scales of “muddy/clear,” “bright/dark,” “smooth/rough,” and “strong/weak,” as shown in Table 8. The rank correlation coefficients between the numbers of voiced and voiceless consonants and the mean scores on the “powerful/powerless” scale were also statistically significant at \( p < 0.05 \) (voiced consonant—“powerful/powerless”: \(-0.576, d.f. = 14, t = 2.64\); voiceless consonant—“powerful/powerless”: \(0.533, d.f. = 14, t = 2.36\)). These correlation coefficients

| Vowel /i/ | “bright/dark” (-0.677, d.f. = 14, t = 3.44) | “sharp/dull” (-0.712, d.f. = 14, t = 3.79) |
| Vowel /o/ | “bright/dark” (0.769, d.f. = 14, t = 4.50) | “sharp/dull” (0.931, d.f. = 14, t = 9.54) |
| Voiced consonant | “muddy/clear” (-0.788, d.f. = 14, t = 4.79) | “bright/dark” (0.814, d.f. = 14, t = 5.24) |
| | “smooth/rough” (0.693, d.f. = 14, t = 3.60) | “strong/weak” (-0.720, d.f. = 14, t = 3.88) |
| Voiceless consonant | “muddy/clear” (0.780, d.f. = 14, t = 4.66) | “bright/dark” (-0.776, d.f. = 14, t = 4.60) |
| | “smooth/rough” (-0.657, d.f. = 14, t = 3.26) | “strong/weak” (0.716, d.f. = 14, t = 3.84) |
indicate that voiced consonants are frequently used to represent stimuli inducing impressions of "muddiness," "darkness," "roughness," "strength," and "powerfulness." Conversely, voiceless consonants are frequently used to represent stimuli inducing impressions of "clearness," "brightness," "smoothness," "weakness," and "powerlessness.

In Experiment 1, the adjective scales of "muddy/clear," "bright/dark," and "smooth/rough" had high loadings for the clearness factor, while those of "strong/weak" and "powerful/powerless" had high loadings for the powerfulness factor (see Table 2). Furthermore, according to the correlation analysis between the phonetic parameters and the auditory impressions in Experiment 1 (see Table 6), voiced consonant correlated with both the clearness and powerfulness factors. Accordingly, the relationships between voiced consonants and the above adjective scales in Experiment 2 are consistent with the results of Experiment 1.

4. SUMMARY

In Experiment 1, psychoacoustic experiments using onomatopoeic representations were carried out for environmental sounds. Onomatopoeic representations were classified based on the similarities of phonetic features using a hierarchical cluster analysis. As a result, similar acoustic properties were found in the stimuli that were expressed by onomatopoeic representations classified into the same clusters. This result suggests the validity of onomatopoeic representations for identifying the acoustic properties of environmental sounds. Furthermore, the auditory impressions associated with sound stimuli were measured by the semantic differential method. The 13 SD scales were integrated into the following three factors by factor analysis: emotion, clearness, and powerfulness. The following relationships among the acoustic properties, onomatopoeic features, and auditory impressions associated with sound stimuli were found. The voiced consonant was often used to represent sounds displaying components in the frequency range below about 1 kHz, inducing impressions of "powerfulness" along with "darkness" and "dullness." Furthermore, the Japanese vowel /i/ was frequently used to represent sounds with spectral centroids at approximately 5 kHz, inducing impressions of "sharpness" and "brightness." Conversely, vowel /o/ was frequently used to represent sounds with spectral centroids at approximately 1.5 kHz, inducing impressions of "dullness" and "darkness."

In Experiment 2, in order to verify the relationships obtained in Experiment 1, psychoacoustic experiments using onomatopoeic representations and measurements of auditory impressions were carried out for synthesized band noises displaying eight kinds of spectral distributions. In Experiment 2, the participants evaluated their impressions of stimuli on adjective scales that had high loadings for the clearness and powerfulness factors in Experiment 1. The results of Experiment 2 confirmed the following relationships obtained in Experiment 1: vowel /i/ was frequently used to represent noise displaying energy mainly in the high-frequency region. Conversely, vowel /o/ was used to represent noise displaying energy mainly in the low-frequency region. Voiced consonants were frequently used for sound displaying energy in the whole frequency range below about 1 kHz. Furthermore, the following relationships were obtained from the results of a correlation analysis between the phonetic parameters in onomatopoeic representations and the mean scores on the adjective scales. Vowel /i/ was frequently used to represent stimuli inducing impressions of "brightness" and "sharpness," while vowel /o/ was frequently used to represent stimuli inducing impressions of "darkness" and "dullness." Furthermore, voiced consonants were used for stimuli inducing impressions of "muddiness," "darkness," "roughness," "strength," and "powerfulness." These relationships between phonetic parameters and auditory impressions were also consistent with the results of Experiment 1. Conversely, voiceless consonants were used for stimuli inducing impressions of "clearness," "brightness," "smoothness," "weakness," and "powerlessness."

In the present study, probable candidates for describing the relationships among the acoustic properties, onomatopoeic features, and auditory impressions associated with sound stimuli were obtained. These results indicate that onomatopoeic representation offers a valid method for identifying the acoustic properties and auditory impressions associated with sounds. Onomatopoeia can therefore be used as a measure of human impressions in place of standard techniques such as the semantic differential method, which uses adjective pairs.

The sensitivities for particular sounds and their onomatopoeic representations are generally thought to be different depending on the culture. Accordingly, the obtained results are not necessarily applicable to other linguistic areas. However, the "sharpness" impression as it relates to the spectral shape of sound, which was actually included in the clearness factor in Experiment 1, is one of the basic factors in timbre. Therefore, the relationships among the acoustic properties, onomatopoeic features, and auditory impressions concerning the "sharpness" impression may be applicable to other linguistic areas.

ACKNOWLEDGEMENTS

The authors would like to thank all of the participants for their participation in the experiments. They also wish to thank Dr. Jonathan Goodacre for proofreading this paper, and Ms. Fumino Obata for her assistance in the analysis. This study was supported by a Grant-in-Aid for Scientific
Research (No. 15300074) from the Ministry of Education, Culture, Sports, Science and Technology; the governmental “Center of Excellence (COE)” program entitled “Design of Artificial Environments on the Basis of Human Sensibility”; and the User Science Institute, Kyushu University.

REFERENCES


