FREE CLASSIFICATION OF VOCAL IMITATIONS OF EVERYDAY SOUNDS

Arnaud Dessein, Guillaume Lemaitre
dessein@ircam.fr, lemaitre@ircam.fr

IRCAM – 1, place Igor Stravinsky – 75004 Paris

ABSTRACT

This paper reports on the analysis of a free classification of vocal imitations of everyday sounds. The goal is to highlight the acoustical properties that have allowed the listeners to classify these imitations into categories that are closely related to the categories of the imitated sound sources. We present several specific techniques that have been developed to this end. First, the descriptions provided by the participants suggest that they have used different kinds of similarities to group together the imitations. A method to assess the individual strategies is therefore proposed and allows to detect an outlier participant. Second, the participants' classifications are submitted to a hierarchical clustering analysis, and clusters are created using the inconsistency coefficient, rather than the height of fusion. The relevance of the clusters is discussed and seven of them are chosen for further analysis. These clusters are predicted perfectly with a few pertinent acoustic descriptors, and using very simple binary decision rules. This suggests that the acoustic similarities overlap with the similarities used by the participants to perform the classification. However, several issues need to be considered to extend these results to the imitated sounds.

1 INTRODUCTION

1.1 Framework

Vocal imitations are very commonly and spontaneously used in everyday conversations when trying to describe a sound. Two kinds of imitations have to be distinguished: standardized imitations (i.e. onomatopoeias) and non-standardized ones. Onomatopoeias are words, the spelling of which is conventional, and the meaning shared by a given population. "Cock-a-doodle-doo" is an example of onomatopoeia in English: every English listener knows that this word labels the cry of a rooster, but its pronunciation might be somehow different from the rooster's cry. On the contrary, non-standardized imitations occur when a speaker tries to imitate a sound with any means of vocal production, without using standardized words. Whereas there is a finite number of onomatopoeias in a language, the variety of vocal imitations potentially occurring in conversations is virtually infinite.

The study reported here focuses on non-standardized vocal imitations. The assumption is made that such imitations are simplifications of the imitated sounds, which still allow the recognition of what has been imitated. Therefore, the production of vocal imitations is believed to provide a relevant paradigm for the study of how human listeners identify sound sources. Sound source identification and the perception of everyday sounds have become an important domain of research since the 90's [6, 7], the potential applications of which are manyfold in audio content analysis or sound synthesis. More specifically, studying sound source identification and vocal imitations is expected to inform the development of *cartoonification*, a particular method of sound synthesis that consists in exaggerating some acoustic features while discarding some others [18]. The advantages of such a technique are that it renders the information clearer, more effective, while reducing the computational cost.

1.2 State-of-the-art

Vocal imitations have been studied from different perspectives. Laas et al. [11,12] showed that listeners could identify fairly well human-imitated animal sounds, and that the identification performances were sometimes even better with imitations than with real animal sounds. Nevertheless, the authors do not explicitly mention whether participants listened to the sounds to imitate or were given the names of the animals to imitate, and whether they could used onomatopoeias or not. Therefore, the successful identification might be accounted for the conventionality and symbolism in the imitations. Other studies have reported systematic patterns of associations between phonetic properties of the imitations and acoustical properties of the imitated sounds [2,4,21,22]. For example, plosives are very commonly used to imitate short sounds or sounds with brutal onsets, such as impacts, explosions. Fricatives are used to imitate sounds with smooth onsets, such as the wind, a breath. The length of imitations is related to the duration of the sounds, or to the number of distinct elements composing the sounds. This shows that vocal imitations can mimic various temporal aspects

SMC 2009, July 23-25, Porto, Portugal

of sounds. However, there are always imitations that do not verify these rules, and some imitations are better than others. Other studies were interested in vocal imitations of tabla drumming sounds [15], strange machine sounds [14, 20], impulse sounds [8, 9], various sounds [10], flue organ pipe sounds [17], sounds of laser printers and copy machines [19]. They show that many spectro-temporal properties of the sounds are reproduced in the imitations: duration, range of frequency, spectral centroid, transients, etc.

Overall, this review of the literature points out several issues. Not all imitations allow perfect identification. The quality of an imitation can be related to the capacities of the vocal apparatus, the performance of the imitators, the difficulty to imitate a given sound. The degree of conventionality and symbolism of the imitations is variable. This can be linked with the nature of the imitations (onomatopoeias, non-word phonetic imitations or non-phonetic imitations).

1.3 Outlines of the study

The analysis reported in this article is based on the results of an experimental study described in [3]. This study provides a set of vocal imitations of everyday sounds that have been categorized by a group of listeners. These imitations are non-standardized, and some of them are even difficult to transcribe phonetically. The goal of this paper is to highlight the acoustical properties that have allowed the listeners to classify the imitations into categories that are closely related to the categories of the imitated sound sources. This paper reports on the specific techniques that have been developed to this end, as well as the results of the analyses.

The experimental study is reported in Section 2. The participants' strategies are analyzed in Section 3, with a specific technique to detect the outliers, using the R_V coefficient. The categories provided by the participants are analyzed in Section 4, with hierarchical clustering and the inconsistency coefficient. The acoustical properties accounting for the categories of imitations are finally highlighted in Section 5.

2 FREE CLASSIFICATION OF IMITATIONS

2.1 Recordings

Vocal imitations were recorded for a set of environmental sounds. The imitated sounds were selected from a corpus of sounds recorded in a kitchen. These sounds had already been used in other experimental studies reported in [13]. Particularly, they had been used in a free classification task. Therefore, the perceptual organization of these sounds into categories of sound sources is available (the 4 main categories are liquid, solid, gas, electric). During the recording session, the participants listened to each sound to imitate and had three trials to record an imitation. They were explicitly asked not to use words, in particular onomatopoeias.

2.2 Method

Twelve sounds were chosen: 3 liquid sounds L_1, L_2, L_3, 3 solid sounds S_1, S_2, S_3, 3 gas sounds G_1, G_2, G_3, and 3 electric sounds E_1, E_2, E_3. Six imitators were chosen: 3 women W_1, W_2, W_3, and 3 men M_1, M_2, M_3. The 6 imitations of each of the 12 sounds gave a corpus of $n = 72$ vocal imitations that were used in a free classification experiment. Participants had first to group together the imitations so as to form different classes. They could create as many classes as they wished, and did not receive any specific instruction on how to form the classes. Then, they had to freely describe each class they had made. For each participant p, the results of the classification were encoded in a $n \times n$ matrix $\mathbf{D}_p$, called *distance matrix*, such that:

$$d_{ij}^{(p)} = \begin{cases} 0 & \text{if sounds } i \text{ and } j \text{ were grouped together;} \\ 1 & \text{else.} \end{cases} \quad (1)$$

3 THE PARTICIPANTS' STRATEGIES

3.1 Descriptions of the categories

The descriptions of the categories provided by the participants are not systematically analyzed here. They suggest however that the participants have used different kinds of similarities to group together the sounds (according to the typology defined in [13]). Indeed, most of the descriptions mention *causal* and *semantic* similarities (i.e. the cause and the meaning associated with the identified sources). But other similarities were also used: acoustical properties of the sounds, feelings (called here *hedonic* properties), means of vocal production (see Table 1). For some participants, the description of a given class sometimes mentions several kinds of similarities (e.g. "Continuous sounds, with a kind of vibration, with the lips, the throat, there is something spinning, noises of machines"). Furthermore the descriptions provided by a participant suggest that he made classes in a rather random fashion.

Similarity	Examples of descriptions
Causal / Semantic	"Mechanical actions of slicing" "Water dripping" "All kinds of drilling machines, food processors"
Acoustic	"Loud and rhythmic sounds" "Repeated, percussive sounds" "Continuous sounds"
Hedonic	"Very aggressive, catches attention" "Suffering" "Mentions the comfort"
Vocal production	"Throat noises" "Expiration with a whistle on the tongue" "With the lips"

Table 1. Examples of descriptions given by the participants, sorted into different kinds of similarities.

3.2 Individual classifications

The descriptions of the classes suggest different strategies across the participants, and even an outlier behaving randomly. There is however no widespread method to analyze individual differences in classification experiments. We used here a method inspired from [1]. It consists in computing a measure of pairwise similarity between the individual classifications. It is also possible to add random individual classifications in order to detect potential outliers.

3.2.1 A measure of pairwise similarity

The R_V coefficient [5] is a measure of similarity between two symmetric matrices $\mathbf{X}$ and $\mathbf{Y}$ and is given by:

$$R_V(\mathbf{X},\mathbf{Y}) = \frac{\operatorname{trace}(\mathbf{X}\mathbf{Y}^T)}{\sqrt{\operatorname{trace}(\mathbf{X}\mathbf{X}^T)\operatorname{trace}(\mathbf{Y}\mathbf{Y}^T)}} \qquad (2)$$

Therefore, it can be used as a measure of pairwise similarity between individual classifications. Following [1], the R_V coefficient is not computed here directly between the distance matrices, but between the individual normalized (with respect to the spectral radius) *cross-product matrices*. The cross-product matrix $\tilde{\mathbf{S}}_p$ for participant p is given by:

$$\tilde{\mathbf{S}}_p = -\frac{1}{2}\,\mathbf{C}\mathbf{D}_p\mathbf{C}^T \qquad (3)$$

where $\mathbf{D}_p$ is the distance matrix of participant p. The $n \times n$ matrix $\mathbf{C}$ is called a *centering matrix* and is given by:

$$\mathbf{C} = \mathbf{I} - \mathbf{1}\cdot\mathbf{m}^T \qquad (4)$$

where $\mathbf{I}$ is the $n \times n$ identity matrix, $\mathbf{1}$ is a column vector of length n filled with ones, and $\mathbf{m}$ a column vector of length n called *mass vector* and composed of positive numbers whose sum is equal to 1. Here, all observations are of equal importance so we set each element of $\mathbf{m}$ equal to $\frac{1}{n}$.

3.2.2 Distances between participants

The *between-participant similarity matrix* $\mathbf{R_V}$, whose coefficients $[\mathbf{R_V}]_{ij} = R_V(\mathbf{S}_i,\mathbf{S}_j)$ are the R_V coefficients between the normalized cross-product matrices $\mathbf{S}_i$ and $\mathbf{S}_j$, is then constructed. A principal component analysis (PCA) is applied on $\mathbf{R_V}$ and is represented in Figure 1 using the two principal components. In this map, a kind of distance between the participants is represented since the proximity between two points reflects their similarity. We also added random normalized cross product matrices in order to simulate random individual results. The participant P_{13} that we suspected to be an outlier is closer to the random participants than the other real participants. He was therefore excluded for the rest of the analysis. However, the presented technique did not allow to highlight different strategies across the participants, even when using more dimensions and removing the random individual results.

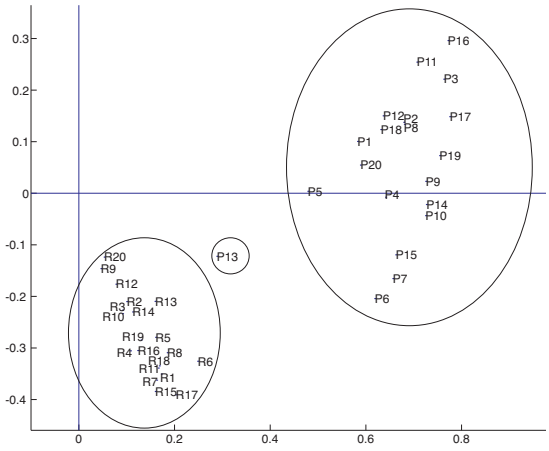

Figure 1. Representation of the distances between the participants using the two principal components of the PCA applied on the between-participant similarity matrix $\mathbf{R_V}$, with the real participants (P) and random participants (R).

4 ANALYSIS OF THE CLASSIFICATION

4.1 Hierarchical clustering

The average distance matrix $\mathbf{D}$ across the individual matrices $\mathbf{D}_p$ was submitted to a *hierarchical clustering* analysis, which represents the average distances in $\mathbf{D}$ with a tree called *dendrogram*. In this tree, the distance between two items (here vocal imitations) is represented by their *height of fusion* (i.e. the height of the node linking the two items).

To identify significant clusters of items, the dendrogram is usually cut at a given height of fusion. As an alternative clustering method, we propose here to use a threshold of *inconsistency*. The advantage of the inconsistency is to emphasize compact subclasses that would not be revealed using the height of fusion. The inconsistency coefficient characterizes a given node by comparing its height of fusion with the respective heights of fusion of its non-leaf subnodes:

$$\text{inconsistency} = \frac{\text{height of fusion} - \mu_d}{\sigma_d} \qquad (5)$$

where μ_d and σ_d are respectively the mean and the standard deviation of the height of fusion of the d highest non-leaf subnodes. The depth d specifies the maximum number of non-leaf subnodes to include in the calculation. The maximum number is used if there are enough non-leaf subnodes, otherwise all non-leaf subnodes are included. The inconsistency coefficient of a given node is positive, having a value set to 0 for leaf nodes, and increasing with the inner dissimilarity of the objects merged by that node.

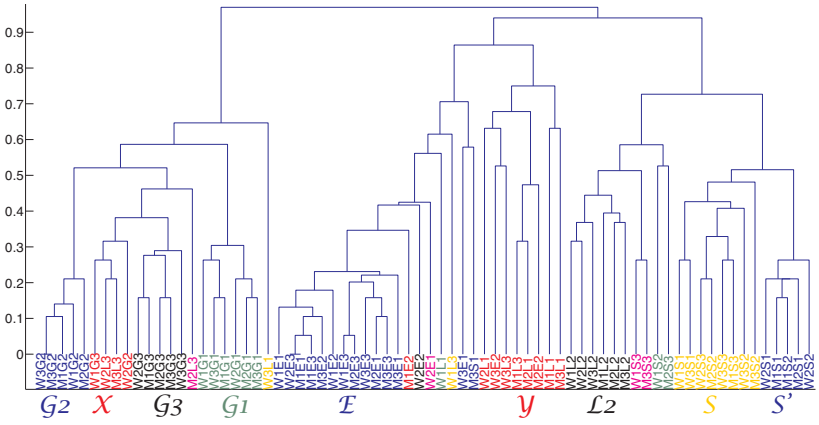

Figure 2. Dendrogram of the vocal imitations (labeled using the imitator's label followed by the label of the imitated sound).

4.2 Dendrogram of the vocal imitations

The dendrogram of the imitations is represented in Figure 2 (using the unweighted average linkage method). We created the clusters with a threshold of inconsistency equal to 1.45 (using a maximal depth so that for each node, all its non-leaf subnodes are included in the calculation). We chose this threshold by decreasing the inconsistency, and so increasing the number of clusters, until the created clusters did not seem coherent to us anymore. It is important to note that the 6 imitations of a given sound are not systematically in the same cluster — in fact, only the sounds G_1 and L_2 have their 6 imitations clustered together. Our hypothesis is that it is related to the quality of the vocal imitations of a given sound, or at least to the agreement between participants on the way to imitate a given sound. As we want to highlight common acoustic invariants in the imitated sounds, we focused on 7 clusters that seemed relevant to us:

(1) $\mathcal{G}_1$ made up of 6 imitations of the gas G_1;

(2) $\mathcal{G}_2$ made up of 5 imitations of the gas G_2;

(3) $\mathcal{G}_3$ made up of 5 imitations of the gas G_3;

(4) $\mathcal{E}$ made up of 12 imitations of electric sounds;

(5) $\mathcal{L}_2$ made up of 6 imitations of the liquid L_2;

(6) $\mathcal{S}$ made up of 8 imitations of solids;

(7) $\mathcal{S}'$ made up of 5 imitations of solids.

We rejected $\mathcal{X}$ because it contains 2 imitations of a liquid and 1 imitation of two gases. We also rejected $\mathcal{Y}$ because although it contains 4 imitations of the same liquid L_1, it also contains 2 imitations of another liquid and 2 imitations of an electric sound, and because its node of fusion is quite high. Finally, we did not consider the clusters with 1 or 2 imitations (the other clusters gather at least 4 imitations).

5 ACOUSTIC PROPERTIES OF THE IMITATIONS

The goal of the analysis reported in this section is to predict the classification of the 7 clusters described above from the acoustical properties of the sounds. We used binary decision trees with a few relevant acoustic descriptors. The descriptors were computed with the IrcamDescriptor toolbox [16].

5.1 First level of the hierarchy

As a first step, we considered the first 3 classes in term of height of fusion: $\mathcal{G}$ composed of $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$, $\mathcal{E}$ as described previously, and $\mathcal{R}$ composed of $\mathcal{L}_2$, $\mathcal{S}$ and $\mathcal{S}'$. To explain these classes, we chose 2 descriptors: (MA) the modulation amplitude of the energy envelope to discriminate between the sounds with a repetitive pattern in $\mathcal{R}$ and the one-block sounds in $\mathcal{G}$ and $\mathcal{E}$, and (MSC) the loudness weighted mean of the perceptual spectral centroid to discriminate between the unvoiced imitations with a high-frequency noisy part in $\mathcal{G}$ and the voiced imitations with a relatively low fundamental frequency in $\mathcal{E}$. The classes are perfectly discriminated (see Figure 3) with the following rules:

(1) $\mathcal{G}$: MA < 0.301208 and MSC ≥ -0.0697998;

(2) $\mathcal{E}$: MA < 0.301208 and MSC < -0.0697998;

(3) $\mathcal{R}$: MA ≥ 0.301208.

One may wonder if these rules generalize well if considering the first 3 classes with the 72 imitations, instead of the first 3 classes with the 47 imitations of the 7 clusters. The answer is rather positive even if there are 7 errors of classification (see Figure 4). These errors are in part due to the fact that 2 of the 3 classes have imitations with a repetitive pattern, whereas only $\mathcal{R}$ has such imitations within the 7 clusters considered. Thus MA is not sufficient anymore to discriminate between 2 of the 3 classes.

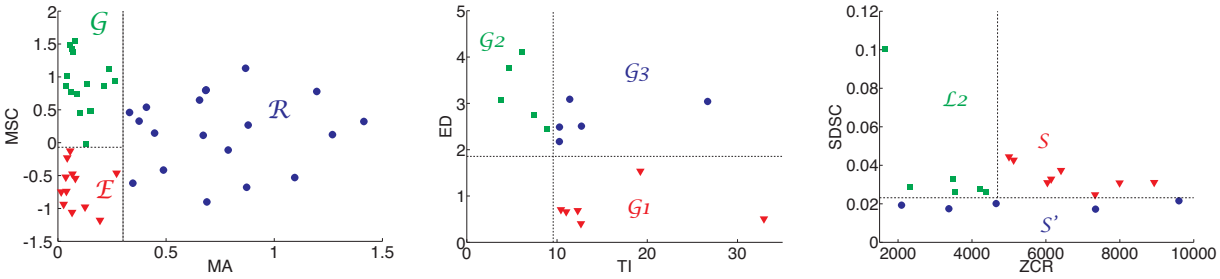

Figure 3. Discrimination between (1) $\mathcal{G}$, $\mathcal{E}$, $\mathcal{R}$, (2) $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3 \subset \mathcal{G}$, and (3) $\mathcal{L}_2$, $\mathcal{S}$, $\mathcal{S}' \subset \mathcal{R}$ (from left to right) with binary decision rules and a few relevant acoustic descriptors.

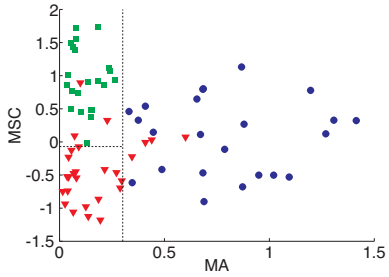

Figure 4. Generalization of the binary decision rules for the discrimination between $\mathcal{G}$, $\mathcal{E}$ and $\mathcal{R}$, to the discrimination between the first 3 classes with the 72 vocal imitations.

5.2 Second level of the hierarchy

We then focused on $\mathcal{G}$ and explained the 3 subclasses $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ with 2 descriptors: (TI) the temporal increase of the energy envelope, and (ED) the effective duration of the energy envelope, because we remarked that the sounds in $\mathcal{G}_1$ are short with a brutal attack, the sounds in $\mathcal{G}_2$ are long with a smooth attack, and the sounds in $\mathcal{G}_3$ are long with a brutal attack. The classes are perfectly discriminated (see Figure 3) with the following rules:

(1) $\mathcal{G}_1$: ED < 1.85578;

(2) $\mathcal{G}_2$: ED ≥ 1.85578 and TI < 9.57589;

(3) $\mathcal{G}_3$: ED ≥ 1.85578 and TI ≥ 9.57589.

We also focused on $\mathcal{R}$ and explained the 3 subclasses $\mathcal{L}_2$, $\mathcal{S}$ and $\mathcal{S}'$ with 2 descriptors: (SDSC) the loudness weighted standard deviation of the perceptual spectral centroid to discriminate between the sounds with a varying timbre in $\mathcal{L}_2$ and $\mathcal{S}$ and the sounds with a constant timbre in $\mathcal{S}'$, and (ZCR) the loudness weighted mean of the zero-crossing rate to discriminate between the sounds with a quite low pitch in $\mathcal{L}_2$ and the sounds with a higher pitch and some kind of noise and roughness in $\mathcal{S}$. The classes are perfectly discriminated (see Figure 3) with the following rules:

(1) $\mathcal{L}_2$: SDSC ≥ 0.0231424 and ZCR < 4692.32;

(2) $\mathcal{S}$: SDSC ≥ 0.0231424 and ZCR ≥ 4692.32;

(3) $\mathcal{S}'$: SDSC < 0.0231424.

6 DISCUSSION AND PERSPECTIVES

This paper reports on the analysis of a free classification experiment with vocal imitations of environmental sounds. The descriptions provided by the participants suggest that they used different kinds of similarities to group together the imitations: causal, semantic, acoustic, hedonic, types of vocal production. We have therefore proposed a method to assess the individual strategies. Using the R_V coefficient, we computed a measure of pairwise similarity between the participants. Although we detected an outlier, we were not able to highlight different strategies. This may be due to the method, or to the fact that the different kinds of similarities might actually overlap. This method must therefore be tested on synthetic data and other results from classification experiments to assess its robustness and reliability. Other measures of pairwise similarity could alternatively be used.

The participants' classifications were submitted to a hierarchical clustering analysis. We created clusters using the inconsistency coefficient, instead of the height of fusion. We chose a relevant threshold of inconsistency and created 7 clusters, which seemed interesting for finding acoustic invariants involved in the recognition of the imitated sources. However, a more systematic method to select the threshold of inconsistency may be preferred. A potential technique based on bootstrap is currently being developed.

It was finally possible to predict the 7 clusters by using binary decision rules with a few acoustic descriptors. With only 6 relevant descriptors, we discriminated the clusters perfectly. This suggests that the acoustic similarities

overlap with the similarities used by the participants to perform the classification. Several issues need to be considered to extend these results to the imitated sounds. We worked with non-onomatopoeic imitations so as to emphasize their acoustic properties, and we chose clusters with respect to their relative quality. But we should now assess the quality of the imitations and their symbolic aspect, to ensure that the acoustic invariants found in the imitations can be generalized to the real sounds. Indeed, the imitations may allow the discrimination between the perceptual categories but not the recognition of these classes. Further experiments are required to address this issue.

7 ACKNOWLEDGMENTS

This work was funded by the EU project *CLOSED* FP6-NEST-PATH no. 29085.

8 REFERENCES

[1] H. Abdi, D. Valentine, S. Chollet, and C. Chrea. Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications. *Food quality and preference*, 18(4):627–640, 2002.

[2] A. Akiyama. Analysis of onomatopoeia in French. Undergraduate thesis, Tokyo University of Foreign Studies, Tokyo, Japan, 2001.

[3] K. Aura. Imitation et catégorisation des sons dans le développement normal [Sound imitation and categorization in normal development]. Master's thesis, Université de Toulouse le Mirail, Toulouse, France, 2007.

[4] R. A. W. Bladon. Approaching onomatopoeia. *Archivum Linguisticum Leeds*, 8(2):158–166, 1977.

[5] Y. Escoufier. Le traitement des variables vectorielles [The treatment of vector variables]. *Biometrics*, 29:751–760, 1973.

[6] W. W. Gaver. How do we hear in the world? Explorations in ecological acoustics. *Ecological Psychology*, 5(4):285–313, 1993.

[7] W. W. Gaver. What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993.

[8] K. Hiyane, N. Sawabe, and J. Iio. Impulse sound recognition system based on onomatopoeia. In *Proceedings of Acoustical Society of Japan*, pages 135–136, 1998.

[9] J. Iio and K. Hiyane. Onomatopoeia cluster for non-speech recognition. In *Proceedings of IEICE*, 1999.

[10] S. Iwamiya and M. Nakagawa. Classification of audio signals using onomatopoeia. *Journal of Soundscape Association of Japan*, 2000.

[11] N. J. Laas, S. K. Eastham, T. L. Wright, A. R. Hinzman, K. J. Mills, and A. L. Hefferin. Listeners' identification of human-imitated animal sounds. *Perceptual and Motor Skills*, 57:995–998, 1983.

[12] N. J. Laas, A. R. Hinzman, S. K. Eastham, T. L. Wright, K. J. Mills, B. S. Bartlett, and P. A. Summers. Listeners' discrimination of real and human-imitated animal sounds. *Perceptual and Motor Skills*, 58(2):453–454, April 1984.

[13] G. Lemaitre, O. Houix, N. Misdariis, and P. Susini. Listener expertise and sound identification influence the categorization of environmental sounds. Submission to *Journal of Experimental Psychology: Applied*, 2009.

[14] M. Ono, T. Sato, and K. Tanaka. Basic study on applying onomatopoeia to evaluating strange sound (Second report: Study on uttered sound of onomatopoeia). *Symposium on Evaluation and Diagnosis*, 3:29–32, 2004.

[15] A. D. Patel and J. R. Iversen. Acoustic and perceptual comparison of speech and drum sounds in the north indian tabla tradition: an empirical study of sound symbolism. In *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, August 2003.

[16] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, IRCAM, 2004.

[17] V. Rioux. *Sound quality of flue organ pipes. An interdisciplinary study on the art of voicing*. PhD thesis, Chalmers University of Technology, Sweden, 2001.

[18] D. Rocchesso, R. Bresin, and M. Fernström. Sounding objects. *IEEE Multimedia*, 10(2):42–52, 2003.

[19] M. Takada, K. Tanaka, S. Iwamiya, K. Kawahara, A. Takanashi, and A. Mori. Onomatopoeic features of sounds emitted from laser printers and copy machines and their contribution to product image. *Journal of INCE/J*, 26:264–272, 2002.

[20] K. Tanaka. Study of onomatopoeia expressing strange sounds (case if impulse sounds and beat sounds). *Transactions of the Japan Society of Mechanical Engineers*, 1995.

[21] H. Wissemann. *Untersuchungen zur Onomatopoiie [Study on onomatopoeia]*. Winter Verlag, 1954.

[22] R. Zuchowski. Stops and other sound-symbolic devices expressing the relative length of referent sounds in onomatopoeia. *Studia Anglica Posnaniensia*, 33:475–485, 1998.