Audio Engineering Society

Convention Paper

Presented at the 126th Convention
2009 May 7–10 Munich, Germany

Generic Sound Effects to Aid in Audio Retrieval

David Black[1], Sebastian Heise[1], and Jörn Loviscach[2]

[1 1]Hochschule Bremen (University of Applied Sciences)
dblack@stud.hs-bremen.de, sebastian@h3e.eu
[2]Fachhochschule Bielefeld (University of Applied Sciences)
jl@j3L7h.de

ABSTRACT

Sound design applications are often hampered because the sound engineer must either produce new sounds using physical objects, or search through a database of sounds to find a suitable sample. We created a set of basic sounds to mimic these physical sound-producing objects, leveraging the mind's onomatopoetic clustering capabilities. These sounds, grouped into onomatopoetic categories, aid the sound designer in music information retrieval (MIR) and sound categorization applications. Initial testing regarding the grouping of individual sounds into groups based on similarity has shown that participants tended to group certain sounds together, often reflecting the groupings our team constructed.

1. INTRODUCTION

A sound designer has a limited set of physical sound-producing tools at his disposal with which to record and achieve a desired output. These tools, familiar to Foley, theater, and radio play sound artists, include such items as doorknobs, broken glass, and coconut shells. Although these tools frequently suffice for many applications, due to space, expense, and time considerations, sound designers often turn to recorded sound sample libraries, either their own or prepackaged compilations, to complement their physical tools. However, as sound sample libraries become increasingly larger, the task of browsing and searching these libraries becomes progressively burdensome.

To overcome the difficulty of browsing and searching through sound sample libraries, we present a toolbox of primary sound elements, called "atomic sounds," intended for use in various applications. These include synthesizing new sounds, creating auditory icons, indexing, and finding similar sounds. Following the idea of a Foley artist's toolbox, we searched for sounds that replicate comparable objects and actions. To leverage the Foley artist's mind model, all sounds have to be clearly identified by their physical counterparts.

Our primary tested application is music information retrieval (MIR). In combination with our own similarity-based audio search system [6], the atomic sound toolset will be used by both professional sound designers as well as amateurs who may publish content on websites such as YouTube to synthesize a "query sample" based on the users' preconceived notions of a similar sound that they would like to retrieve from their sound sample library. To create this synthesized query sample, the user would simply select samples from the atomic sound toolbox, arrange them onto a canvas in our paintbrush-stroke-based similarity search tool [5] until they are satisfied with the result, and query the system for similar sounds to their synthesized sample.

The similarity-based sound search tool necessitated a primary set of sounds to use with the included paint-style tools. This collection of sound samples has been created to mimic a set of some of the various physical sound-producing items that are common in sound designers' studios. Following this physical object metaphor, we assume the user might have a close connection to those physical objects, which in turn would help him or her to conceivably search for any related sound. Combining the sounds is similar to recording and arranging sounds of the physical objects.

In order to create a comprehensive set of atomic sounds, libraries of pre-packaged Foley and sound design samples were manually analyzed with the objective of discovering which sounds were fundamentally simple, not composed of two or more identifiable components. As previous research in sound perception has shown, listeners are often able categorize sounds into onomatopoetic groups [9] or according to mental imagery [7][8] when given a large set of sounds. To mimic this mind model, eight onomatopoetic groups were created and suitable example atomic sounds were found to complete each of the groups.

To ascertain the functionality of our atomic sounds, we conducted an evaluations of our sound set in which users were asked to group a selection of random sounds into different categories based on similarity in order to determine whether or not the sounds that we have selected were grouped in similar ways to ours, or rather if users grouped them in ways similar to each other.

2. RELATED WORK

In a previous work, Scavone et al. [7] obtained similarity ratings from human listeners for several hundred sounds in order to train an automated computer audio classifier. A two-dimensional graphics-based software program for collecting similarity data for large sets of sound stimuli was developed. With a drag-and-drop two-dimensional palette, participants created categories and assigned color labels to them. Confidence ratings helped decide which sounds didn't "fit" as well into the groupings. The team employed sound effects from commercial libraries. Sounds were chosen that were producible by humans rather than abstract sounds. The participants "neatly clustered" the sounds according to the participants' mental imagery of the sources of the sounds. Participants encountered difficulties grouping the sounds based on timbre rather than on pitch or loudness, and mentioned that pairwise comparisons were an unintuitive and artificial way to compare sounds, instead preferring the two-dimensional method.

Bonebright [1] described a three-dimensional MDS perceptual structure for a large set of 74 everyday sounds. Relationships between the sounds were tested using both perceptual and acoustic data. She attempted to discover which physical characteristics of the stimuli drove the perceptual process of the participant's listening. Stimuli consisted of 74 sounds produced by objects that a person would encounter on an everyday basis. Using sorting tasks for audio stimuli was found to be effective (as opposed to direct comparison tasks) and correlations between audio attributes were made.

A further study by Stepanek was made [8] to determine lexical dimensions of timbre. Musicians were asked to imagine an orthogonal space of various parameters into which sounds produced by musical instruments could be placed. Four basic dimensions of timbre were found: 1. gloomy — clear 2. harsh — delicate, 3. full — narrow 4. noisy / rustle — non-noisy.

Sundaram et al. [9] classified sound clips based on both semantic and onomatopoetic tags. Onomatopoetic and semantic tags were manually labeled by test subjects, i.e. "completely based on subjective perception". An automatic clustering using feature-vectors yielded clustering accuracy of 60% using a selection of sounds from the BBC sound library placed manually into 20 categories. Using both high-level semantic labels and mid-level onomatopoetic labels provided flexibility, as each scheme counteracted the shortcomings of the other. However, the team noted [10] that inconsistencies in interpreting onomatopoetic labels resulted in some groups of these labels being more separable than others.

Rather than creating their own set of descriptors, Cano et al. [2] employed a semantic framework to help group sounds. They argued that manual audio filing is both error-prone and labor-intensive, because languages are imprecise and informal, and that automatic annotation schemes for sound are not mature [3]. The MPEG-7 framework was used on top of the existing WordNet lexical network to produce a classification scheme that simplifies a librarian's work by allowing for an unambiguous way to link sound effect terms.

3. APPLICATION AND CONCEPT

The impetus for creating and evaluating a basic set of atomic sounds is the proposed use in SonoSketch, a software program developed for use in searching large databases of sound samples.

3.1. SonoSketch

SonoSketch [5] is a program that allows users to sketch query sounds for entry in a sound search program [6] that employs Mel-Frequency Cepstral Coefficients (MFCC) to assist in locating similar sounds in sound effects databases. The user literally sketches this query sound by placing curved strokes and granular cloud shapes on a two-dimensional canvas. The atomic sound set described in this paper comprises the toolset of sounds that are to be represented by the strokes placed by the user on the canvas. To provide the SonoSketch user with a wide, yet compact array of options, the following method was conducted to arrive at this toolbox of atomic sounds.

3.2. Generating the Atomic Sound Set

A basic set of atomic sound samples was sought to provide a sonic toolbox for the SonoSketch application. To create this set of sounds, the contents of the Sound Ideas[1] General 6000 and Hanna Barbera libraries of prerecorded Foley sound samples and sound effects were analyzed. The first part of the analysis involved manually compiling different classes of sounds based on how the sounds were produced or what they represent, such as doors creaking, clothes ripping, or glass breaking. This resulted in a compilation of 65 of the most-occurring classes of sounds. Complex sound samples that included more than one main perceptual sound event were not included. An example of a complex sound would be that of a car crash, which might involve the sounds of a tire screeching, the bang of the collision, and glass shattering.

After these classes were created, multiple onomatopoetic English word tags were manually assigned to each of the classes of sounds, in a process similar to the collection of text descriptors seen in other sound effects similarity schemes [9]. These tags were then used to manually cluster the different semantic classes into similar-sounding categories. This further reduction resulted in eight differently-named categories of onomatopoetic sounds. These categories were then manually filled with between five and ten representative sound samples for a total of 59 samples culled from the Foley sound sample libraries. These sound samples were edited to have lengths of no more than five seconds.

Clunk, Thump, Clink Ping	Crackle, Sizzle, Clatter, Rattle
Creak, Squeak, Break, Crack	Crunch Crush, Crumble, Crash
Drag, Scrape, Scratch, Tear	Slice, Snip, Roll, Slide
Spray, Whirr	Squish, Splat, Stab

Table 1. Onomatopoetic group titles

The resultant selection of onomatopoetic categories (see Table 1) and the sound samples contained therein are the basis for the SonoSketch application, and also constitute the basis for the experiments that were performed to test the hypothesis that users would group these sounds in a similar way when asked to make groupings of the sounds based on similarity.

4. EVALUATION

To better understand the mind model with which a user of the SonoSketch program might approach categorized sounds, and to evaluate the onomatopoetic grouping of our particular selection of sound samples, an experiment concerning the categorized sounds was conducted. This experiment involved users placing a given collection of

[1] http://www.sound-ideas.com/

sounds into groups based on their perceived similarity to others in the group.

4.1. User Test: Setting

An Adobe Flash-based software program was written to enable participants to place 16 randomly-chosen sounds from our selection of 59 into four groups. The software program consists of a top row of 16 play-button-shaped icons, four horizontally-placed square group bins, and buttons for both "abort" and "done". Each play-button icon was able to both play back and cease playback of one of the 16 loaded sounds by either clicking or double-clicking. The "done" button submitted the results, and the "abort" button reset the software so that the user could start again. The "abort" action loaded a new set of 16 randomly-chosen sounds. The test was carried out both in a supervised room 14 times, and on-line 125 times. In total, 139 valid tests were completed.

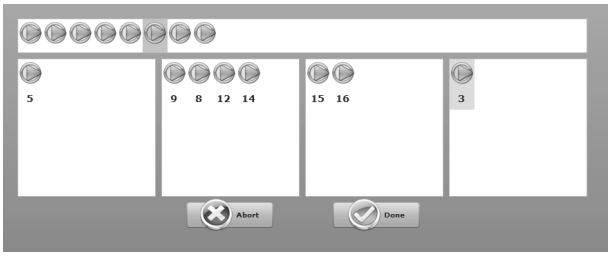

Figure 1: Screenshot of the online survey's user interface, showing the top row of icons and the four bins used to place the grouped sounds.

The participants were asked to drag and drop the 16 play-button icons into the four empty bins so that groups of icons with similar sounds were generated (see Figure 1). The bins could hold any number of the 16 icons. No instructions were given regarding which bins should be used for which sounds, only that they were to hold similar sounds. Sounds could be moved between bins to allow for corrections to be made after they had been initially dragged from the top row. In addition, there was no requirement that all sounds must be used. If a user felt that one particular sound did not match with any others, this could remain in the top row rather than be placed in a bin. After the participant felt that suitable groups had been created, the "done" button was to be clicked.

4.2. User Test: Results

The groups created during the test were recorded in a database table. After completing the test, the groups had been transformed into an adjacency matrix, with the sounds listed by group. We counted how often each file was grouped together with each other one, was grouped together with each other one, correcting for the probability with which each file was presented (see Figure 2).

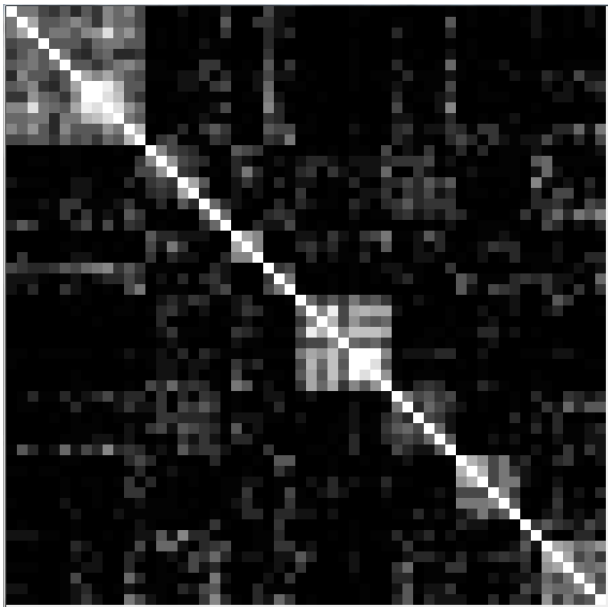

Figure 2: Adjacency matrix according to survey results.

One can see that we found some strong relations between particular files (bright gray and white colors) and some are never grouped together (black color). Clusters and lines of gray and white boxes indicate sounds that tended to be grouped together. The brightness of the gray boxes is related to the number of times the files were grouped together.

We utilized agglomerative hierarchical clustering [4] in which the similarity of two clusters is computed as the geometric mean of all pairwise similarities of their members (in this case, sound files), see Table 2.

1Creak 2Creak 5Creak	1Crunch 1Spray 8Crackle	2Crunch 7Drag 2Drag 5Crunch 7Crackle 6Drag
10Crackle 6Creak 5Crackle 1Crackle 2Crackle	1Slice 5Slice 6Slice	2Slice 2Spray 4Spray 3Slice 3Spray 5Spray
10Clunk 1Clunk 2Clunk 3Clunk 4Clunk 7Clunk 4Crunch 5Clunk 8Clunk 6Clunk 9Clunk 1Drag 4Slice	1Squish 2Squish 3Squish 4Crackle 6Crackle 4Squish 5Squish 9Crackle 3Crackle	3Drag 4Creak 7Creak 4Drag 5Drag 7Slice

Table 2: List of clusters using the geometric mean of each cluster.

To estimate the fitness of our clustering grid we counted how many items must be moved to match our clustering. In Table 3, one can see how many items of a cluster had already fit our model, i.e. 100% of the Clunk sounds were moved into the same cluster by the average user. Most of the creak and crunch sounds were grouped in another way that we assumed. However, 58% of our supposed grid already fits well.

Cluster Fitness	*Onomatopoetic Cluster*
100%	Clunk
30%	Crackle
75%	Slice
100%	Squish
80%	Spray
17%	Creak
43%	Drag
20%	Crunch
58% avg.	

Table 3. fitness between our cluster grid and the measured grid.

In previous work [6] we implemented a self-organizing map (SOM) to build clusters of sound effects in timbre space. From distributing this sound set with the proposed mapping method from SoundTorch we know that this sound set covers large region of the timbral space.

5. CONCLUSION

The atomic sounds outlined in this paper provide a useful model for sound categorization, based on Foley sounds that sound artists use in their routine work. The sounds are useful for applications, such as our SonoSketch program, that require a compact, semantically-related toolbox of short sound effects. The onomatopoetic approach of categorizing sounds lends itself to applications in which users must become familiar with groups of short sound samples. Participants in our test tended to group certain sounds together, and our own grouping of atomic sounds was somewhat supported. Further evaluation must be undertaken into how users group sounds, how they navigate and interact with sets of sounds, and how to achieve the best and most versatile set of basic atomic sounds.

6. REFERENCES

[1] Bonebright, T., "Perceptual structure of everyday sounds: a multidimensional scaling approach," presented at the 7th International Conference on Auditory Display, Espoo, Finland, 2001 July 29-August 1.

[2] Cano, P., Koppenberger, M., Celma, O., Herrera, P., and Tarasov, V., "Sound Effect Taxonomy Management in Production Environments," presented at the AES 25th International Conference, London, United Kingdom, 2004 June 17-19.

[3] Cano, P., Koppenberger, M., Le Groux, S., Ricard, J., Wack, N., and Herrera, P., "Nearest-neighbor Automatic Sound Annotation with a WordNet Taxonomy," J. Intelligent Information Systems, vol. 24, issue 2, pp. 99-111 (2005 May)

[4] Duda, R., Hart, P., Stork, D., Pattern Classification, 2nd Ed., Wiley-Interscience, New York, NY, USA, pp. 552-556. 2001.

[5] Heise, S., Batterman, M., and Loviscach, J., "SonoSketch," presented at the AES 126th Convention, Munich, Germany, 2009 May 7-10.

[6] Heise, S., Hlatky, M., and Lovischach, J., "SoundTorch: Quick Browsing in Large Audio Collections," presented at the AES 125th Convention, San Francisco, United States, 2008 October 2-5.

[7] Scavone, G., Lakatos, S., Cook, P., and Harbke, C., "Perceptual spaces for sound effects obtained with an interactive similarity rating program," presented at the International Symposium on Musical Acoustics, Perugia, Italy, 2001 September 10-14.

[8] Stepanek, J., "Musical Sound Timbre: Verbal Description and Dimensions," presented at the 9th Intl. Conference on Digital Audio Effects, Montreal, Canada, 2006 September 18-20.

[9] Sundaram, S., and Narayanan, S., "Classification of sound clips by two schemes: Using onomatopoeia and semantic labels," presented at the IEEE International Conference on Multimedia & Expo, Hannover, Germany, 2008 June 23-26.

[10] Sundaram, S., and Narayanan, S., "Analysis of Audio Clustering using Word Descriptions," presented at the 32nd IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, United States, 2007 April 15-20.