Audio Engineering Society

Convention Paper

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

SonoSketch: Querying Sound Effect Databases through Painting

Michael Battermann[1], Sebastian Heise[1], and Jörn Loviscach[2]

[1] Hochschule Bremen (University of Applied Sciences), 28199 Bremen, Germany
mbatman@fbe.hs-bremen.de, Sebastian@h3e.eu

[2] Fachhochschule Bielefeld (University of Applied Sciences), 33602 Bielefeld, Germany
jl@j3l7h.de

ABSTRACT

Numerous techniques support finding sounds that are acoustically similar to a given one. It is hard, however, to find a sound to start the similarity search with. Inspired by systems for image search that allow drawing the shape to be found, we address quick input for audio retrieval. In our system, the user literally sketches a sound effect, placing curved strokes on a canvas. Each of these represents one sound from a collection of basic sounds. The audio feedback is interactive, as is the continuous update of the list of retrieval results. The retrieval is based on symbol sequences formed from MFCC data compared with the help of a neural net using an editing distance to allow small temporal changes.

1. INTRODUCTION

Content-based features such as Mel-Frequency Cepstral Coefficients (MFCC) [12] and the low-level audio features from MPEG-7 allow searching for sounds based on acoustic similarity. In a real-world application, however, the "find similar" approach faces a vital issue: Often one does not have a sound at hand to start the search with. We present our solution called SonoSketch to tackle this problem. It facilitates creating a sound that serves as a starting point for a search based on acoustic similarity to find existing recordings in an audio database. The solution is particularly aimed at trained sound designers, for instance Foley artists. They have learned to perceive sound effects in terms of their production: How can other sounds be combined to create the sound that one is looking for? In SonoSketch, the user can choose from a number of generic basic sounds and "draw" with them in a time/pitch coordinate frame. A virtually unlimited number of such components can be layered while still remaining editable, see Fig. 1. This is particularly useful to build composite sounds that develop over time, such as the sophisticated slapstick

sound effects used in the classic cartoon movies of Hanna and Barbera[1].

Fig. 1: Users can "draw" sounds onto a pitch/time grid with different tools.

2. COMPOSING SOUNDS

Observing a Foley artist at work can leave a lasting impression. Professional sound designers use a set of sound tools, such as different floor coverings plus a collection of shoes, miniaturized doors, bags and boxes with stones and broken glass. Although the number of different tools is manageable, the sound effects that can be created with them are beyond imagination. Augmented by some tiny plastic gadgets from common toy stores, these tools are able to produce a wide range of sound effects—extending far beyond door slams and footsteps.

Following this principle, we gathered a set of basic sounds that can be used to create other sounds. However, we aim at a sketch, not at a perfect reproduction. In another work [2] we compiled a collection of sounds effects dedicated for the usage as single sounds in SonoSketch. Furthermore, we provide a solution to quickly select from these basic sounds via a pie menu.

Based on interviews conducted with sound designers, two different kinds of sound events were defined for our tool: first, distinct single sound events that develop time; second, granular, more random sound effects. The prototype offers two different drawing tools to cater for these two categories: a pen and an airbrush (see Fig. 1). The pen can be used for precisely placing a single sound event onto the canvas. The airbrush splatters the surface with random grains.

Depending on the type of tool used to paint on the canvas, different virtual representations of the sounds are displayed: The pen-drawn sounds are rendered as tadpole-like shapes whose tail can be extended arbitrarily and be bent up or down. Granular sounds—placed with the airbrush—are represented through a sponge-like pattern filling a rounded outline defined by a stroke with the mouse.

Whenever the canvas' content change, the resulting sound is synthesized in a background thread and is immediately available for auditioning. Thanks to real-time pitch shifting and time-stretching functions, the sketched sound can be auditioned at any time with no delay that may arise from pre-computation. Furthermore, the sound's acoustic features are extracted and send to the search engine. We aim at a fluid workflow that unifies refining the query and checking the result list.

The software responds to the pressure applied to the tip of a digital pen, if such an input device is used instead of the mouse. We have build one prototype that makes heavy use of the acceleration offered by 3D graphics chips and another prototype that can be used in a Web-based scenario.

Whereas SonoSketch may be used as a graphical audio synthesizer, this is not its major purpose. A vital difference that sets SonoSketch apart from other sound or music drawing tools is that it is intended to create input sounds for content-based Music Information Retrieval (MIR) tasks. Hence, the level of detail in adjustments available to the user is strictly limited and adapted to the task at hand.

To demonstrate its use for audio retrieval, we equipped SonoSketch with MIR functionality. The search for similar sounds is conducted in the background while the user is still creating or refining the sketch. Updated results are presented in parallel to the user's actions in a separated list and are immediately available for auditioning. This seamless integration makes it easy to support the user in refining the sketch. Her or she may emphasize acoustic details that are important for the search, steer the search into promising areas, or turn it away from acoustically close but semantically wrong items (e.g., raindrops as opposed to fireworks).

3. RELATED WORK

This work brings two fields of research together: First, we create a computer interface for visualizing audio

[1] http://hanna-barbera.com

content; second, we combine it with audio content-based retrieval techniques.

Composers such as Arnold Schönberg, Gyorgy Ligeti, and even Jerry Goldsmith experimented with new possibilities of sound synthesis and electronic devices. The harmonic structure of the past dissolved. Thus, the traditional music score reached its limits more and more often.

To deal with the new possibilities, new notations had to be invented for sounds that formerly would have been considered "noise." In the Seventies, for instance, Rainer Wehinger designed a visual listening score to accompany Gyorgy Ligeti's "Artikulation" [18]. He created a graphical language, to describe the different synthesized artificial noises. Round, edgy and comb-styled icons are placed on a grid that represents time and pitch. Colored surfaces group several sound events to build complex relationships.

Countless works have been published in the field of audio visualization, be it for technical analysis or artistic expression. Hyperscore [9][10] is a graphical editor aiming intuitive editing musical structures. Audio and musical features are mapped to different graphical elements and allow high level editing of musical structures as well as it provides access to low level features such as pitch and dynamic. The Sonic Visualizer [3] is another approach for visualizing audio content. It focuses on the visualization of technically extracted audio features of music tracks such as beat and pitch. A tool called Sound Sketch [17] allows the users to draw colored lines onto a grid that represents time and pitch. The purpose of this tool is being a simple synthesizer and not a database query tool. It also does not provide polyphonic sounds or sound effects to draw. U&I Software's MetaSynth[2] can be considered a fully grown implementation of this approach in that it synthesizes sounds from arbitrary images.

The idea of sketching data for retrieval stems from research into query-by-image (QBIC) systems. This forms a major subfield of multimedia retrieval; see Lew et al. [13] for a survey. We bring this approach to music information retrieval (MIR), which is highly active field of research in its own right. Casey et al. [6] survey the known techniques used to analyze audio and extract features that have been successfully used in information retrieval.

[2] http://www.uisoftware.com/MetaSynth/

Many MIR techniques are based on the extraction, tracking, and managing of metadata [8]. State-of-the-art methods use semantic Web technology to weave meta information to meaningful structures. However, using metadata for sound effects is often pushed to the limits. The enormous number of single recordings taken even during a recording session makes it hard to manually keep track of all meta-information. Even worse, there are no known labels for many kinds of useful information. Hence, content-based approaches are inevitable. They operate on the wave alone, possibly extracting metadata in an automated fashion.

Most MIR techniques focus on music classification and not on matching audio snippets and have to be adapted for our purpose. In addition it still is not clear how the extracted features relate to perceived sound similarity. Scavone et al. [16] describe a series of experiments that attempt to merge developments in the study of sound source perception and physical modeling to yield a better understanding of listeners' criteria in rating auditory timbre. The experiments were driven by the need of similarity ratings to train an automated computer classifier.

The computation of timbre similarity based on the acoustic content has been addressed in numerous works. Of major importance to us were these: Kim et al [12] compare the performance of MPEG-7 features and MFCCs for content-based audio retrieval; Casey [5] discusses methods how to calculate similarity for musical data and the different semantics; Peeters [12] presents a method how chroma feature vectors can be successfully used for automatic key estimation in music tracks. audioDB [14] is a database that is specialized for content-based search in vast collections of music tracks. Its feature extraction uses chroma and power values for similarity search. audioDB is a powerful tool for query-by-example—and hence forms a prospective back end of our prototype—, even though it does not provide a method to create a search pattern.

With the increasing speed of desktop computers, real time FFT computations became available to a broad base of users. Hence, real-time pitching and time stretching are well examined. Our implementation is based on a description by Bernsee [7].

4. SKETCHING SOUNDS

To inform our design, we started our research by asking people to sketch images by hand that represent a set of

short audio effects provided by us. The first attempts lead to amusing results. The pictures covered an incredibly wide variety of drawings that range from abstract polygons to dancing penguins, see Fig. 3. In became clear from these early drawings that a common drawing language is hardly to be found by letting let people sketch their ideas without any limitations.

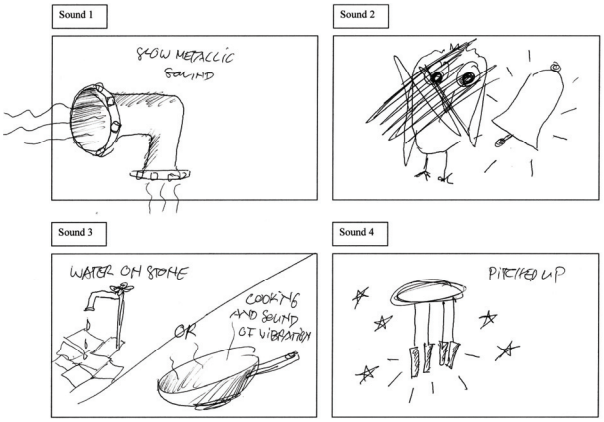

Fig. 2: Participants of the first survey were allowed to freely sketch the sounds they heard.

Many of the sketches we got depict objects that create the corresponding sound. These results are not useable for defining a simplified sketch tool, as we would have to provide a huge collection of sounding items. However, we learned from these results that Foley artists are not the only people to make a close connection between a sound that they hear (or imagine) and a sounding object; they are just trained better, and find it easier to make this connection.

We decided that we need a more abstract and simplified visual language for our purpose and got inspired by Wehinger's drawings (see Fig. 2). However, we expected that the sheer number of symbols and colors offered would still be distracting when visualizing a given sound. Therefore we defined a reduced and simplified version of this particular visual language.

Because the completely free drawings did not produce useful results we conducted a second, more formal experiment in which we limited the graphical options. The subjects were instructed before they started drawing the example sounds. To this end, we presented three sound examples and the corresponding drawings we expected a picture to look like in our proprietary visual language, see Fig. 4. This language consisted of circles and lines on a grid that represents time versus pitch. Circles were to represent a distinct sound event. A line could describe temporal development. For granular sounds, we offered a cloud symbol. This language inherently forces the user to decompose sound effects into separate items on a pitch-and-time grid.

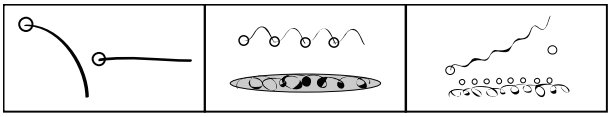

Fig. 3: At the beginning of the second experiment, three examples were shown to the subjects. The corresponding sounds were "Dropping and Ping", "Shaking and Roll" and "Drum-roll and Bend-up".

The 25 subjects (students with digital media and computers science background; 20–35 y) were asked to draw ten sound examples after the introduction described before. The ten sound examples were randomly taken from Sound Ideas's "The General 6000 Edition"[3] which we titled as: "Table Tennis," "Computer Keyboard," "White Noise Doppler Effect," "Jazzy Vibraphone," "Starship Door Opens," "Filling a Glass of Beer," "Arrow Flies and Strikes," "Metallic Chain Drop," "Drilling Machine," and "Washing Machine". One could assume that all of the participants are computer literate and are able to use a graphics tablet. Nonetheless, to avoid a possible bafflement of a subject, we did not use any electronic input device for this experiment, but again asked the subjects to draw with real pens on real paper.

After we applied the introduction to our tests, most of the pictures that different subjects have drawn for the same sound effect became similar. Subjects were able to decompose the example sounds into smaller events and create a visual representation on the pitch-and-time grid from that. Even though we did not define any kind of absolute timeframes or pitch markings on the grid, many users placed the symbols at similar positions, see Fig. 4.

The results indicate that our visual language is a proper technique to describe sound effects. Comparing Fig. 3 with Fig. 1, one can see that the prototype uses graphically enhanced shapes, but that the principle remains the same.

[3] http://www.sound-ideas.com/6000.html

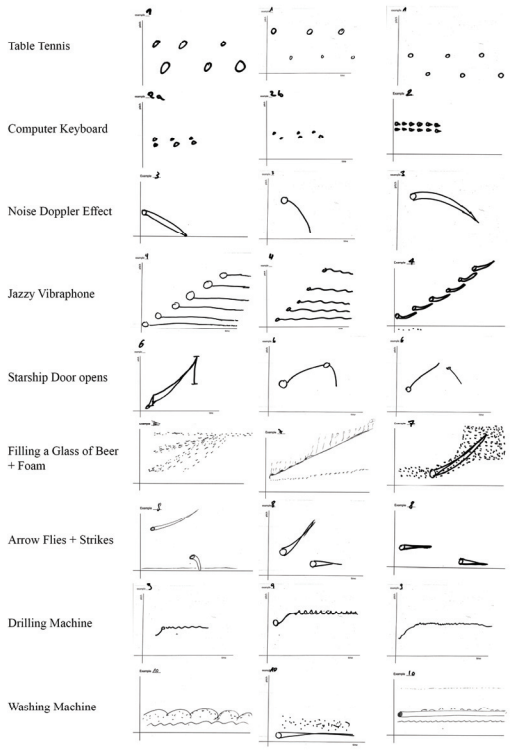

Fig. 4: Sketches from different subjects show that after an introduction to our visual language the sketches become similar.

5. SEARCHING FOR SOUNDS

The second part of the prototype concerns its back end: the database. The symbols drawn onto the surface are synthesized into a sound file using pitch-shifting and time-stretching algorithms. The resulting example is fed into a query-by-example database.

The feature extraction is based on previous work we described in [11]. In a preprocessing step, MFCC features are extracted for frames of 20 ms length with 50 % overlap. The MFCC distance is not sensitive to spectral details, in particular the difference between a metallic sound and noise with the same overall spectral distribution. To enhance the MFCC distance in this respect, we also extract features remotely related to the known chroma vectors.

Because most of our sound effects do take the twelve-tone scale of Western music into account, we used more than the common 12 bins for a chromagram and extended it to 36. The main advantage is that noise still fills all bins uniformly, whereas sounds with a non-uniform spectral pattern such as a bell create a comb-like structure even though this does not have a harmonic structure. Each chroma vector is shifted cyclically so that the highest peak has the index zero. Thus, the harmonic structure can easily be compared through subtracting two vectors. As shown in Fig. 5, the largest peak values (light gray) are in the bottom row. Comparable harmonic structures create the same picture. Fig. 6 shows three different brass bell sounds. Bell 2 and Bell 3 create the same chromagram and would be classified as similar. Bell 1 has a larger distance to the others. However, all bell sounds can be clearly differentiated from a noisy sound.

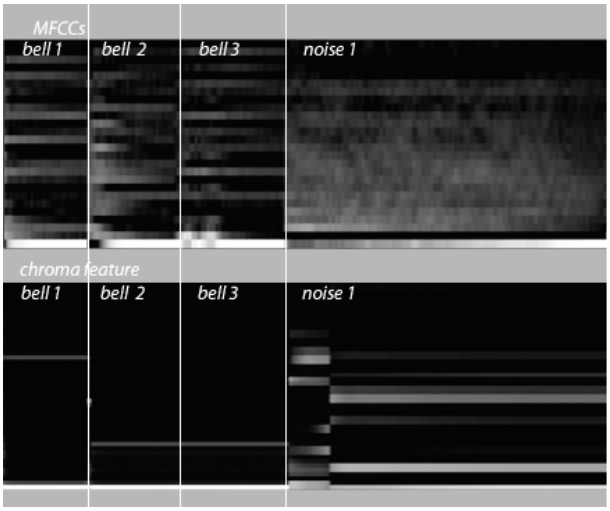

Fig. 5: For the database query, MFCCs (upper row) are combined with chroma-like vectors (lower row).

The query pattern is shifted along the source pattern frame by frame. We compare the query sample with every audio file in the database. The regions with the minimum distance in a source file are tracked and added as an icon to the result window. The result list is updated continuously. Thus, every time the search algorithm finds a better region, it is added to the result list. Only the 50 best results are displayed. To provide a better overview of the results we insert group labels varying from “super fit” to “interesting”.

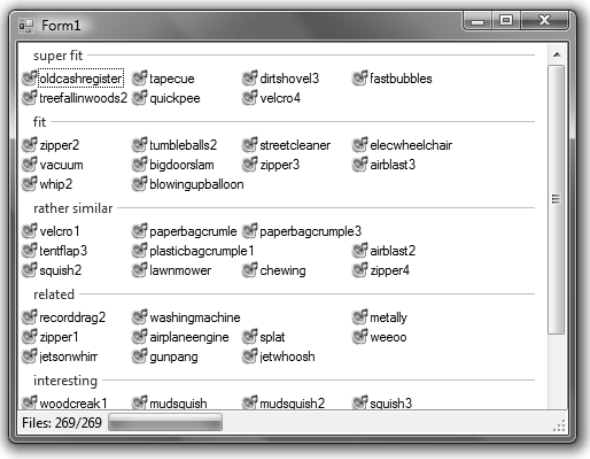

Fig 6: A separate window shows the results of our similarity search, ready to be selected for immediate playback.

We are currently experimenting with neural networks to measure the similarity of audio snippets in future work. We gathered already training data for the net we used an online survey. The participants were asked to estimate the sounding similarity of two audio snippets each of 40 ms length. The extreme short examples avoid conclusions about a possible temporal development leading to semantic interpretation of the audio content. To compile 1000 pairs, snippets have been randomly extracted out of Sound Ideas' "General 6000 Edition." Each pair consists of one reference snippet and a candidate that should be estimated by the subjects. To ensure that there will be some similar candidates we always mix the reference file into the candidate with a randomly taken amount. We provided a simple slider which could be continuously set between "sounds similar" and "total opposite". In each turn we presented 30 pairs to each user. Overall, we received about 2500 results so far.

To generate training data for a neural network from these results we only use the part of the data where the perceived similarity values between all users shows a huge correlation. We found that a neural network can easily be trained with the described feature vectors above as input data to match the data entered by humans. But still we do not have enough statistical data to create a neuronal network that produces robust and satisfying results.

Another issue is indexing the database for faster search. One could try to use clustering algorithms or further abstractions of the feature vectors to create an index table.

6. CONCLUSION

We presented an easy-to-understand tool for querying sound effects from a database. The search—based on query-by-example—is conducted in timbre space and uses audio-only data, based on extracted features. A user can literally sketch a search query onto a canvas by drawing basic sound components with different drawing tools. Changing the basic sound set can adapt SonoSketch not only to search for sound effects but also rhythm patterns or sound samples for music production.

7. REFERENCES

[1] Angalde, A. and Dixon, S., "Characterisation of Harmony with Inductive Logic Programming", presented at the 9th ISMIR Conference, Philadelphia, USA, 2008, September 14-18.

[2] Black, D., Heise, S., "Generic Sound Effects to Aid in Audio Retrieval", AES 126th Convention, Munich, Germany, 2009.

[3] Cannam, C., Landone, C., Sandler, M. and Bello , J.P. "The Sonic Visualiser, A Visualisation Platform for Semantic Descriptors from Musical Signals", in Proceedings of the 7th International Conference on Music Information, Victoria, Canada, October 2006.

[4] Casey, M.A., "AudioDB: Scalable approximate nearest-neighbor search with automatic radius-bounded indexing", 156th ASA Conference, , Miami, USA, 2008, November 10-14.

[5] Casey, M.A., Rhodes, C. and Slaney, M., "Analysis of minimum distances in high-dimensional musical spaces", in IEEE Transaction on Audio, Speech and language processing, Vol. 16, No. 5, 2008.

[6] Casey, M.A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C. and Slaney, M., "Content-based music information retrieval: current directions and Future Challenges", in Proceedings of the IEEE, vol. 96, No. 4, 668–696, 2008.

[7] DSP Dimension:Bernsee, S.M.: Pitch Shifting Using The Fourier Transform. (`http://www.dspdimension.com/admin/pitch-shifting-using-the-ft/`), accessed 2009/01/31.

[8] Fazekas, G., Raimond, Y. and Sandler, M., "A framework for producing rich musical metadata in creative music production", AES 125th Convention, San Francisco, CA, USA, 2008, October 2-5.

[9] Farbood, M., Kaufman, H., Jennings, "Composing with Hyperscore: An intuitive interface for visualizing musical structure.", Proceedings of the International Computer Music Conference, 2007.

[10] Farbood, M., Pasztor, E., Jennings., K., "Hyperscore: A Graphical Sketchpad for Novice Composers. IEEE Computer Graphics and Applications", 2004, January-March.

[11] Heise, S., Hlatky, M., and Loviscach, J., "SoundTorch: Quick Browsing in Large Audio Collections", AES 125th Convention, San Francisco, CA, USA, 2008, October 2-5.

[12] Kim, H.-G.; Moreau, N. and Sikora, T.: MPEG-7 Audio and Beyond - Audio Content Indexing and Retrieval, John Wiley & Sons Ltd, West Sussex, England. 2005

[13] Lew, M.S., Sebe, N., Djeraba, C and Jain, R., "Content-based multimedia information retrieval: State of the art and challenges. In ACM Transactions on Multimedia Computing, Communications, and Applications", Volume 2, pp1-19, 2006.

[14] OMRAS2: audioDB: (`http://www.omras2.org/audioDB`), accessed 2009/02/03.

[15] Peeters, G.: "Chroma-based estimation of musical key from audio-signal analysis", 7th ISMIR Conference, Victoria, Canada, 2006, October 8-12.

[16] Scavone, G.P., Lakatos, S., Cook, P.R. and Harbke, C., "Perceptual Spaces for Sound Effects obtained with an Interactive Similarity Rating Program", presented at the ISMA Conference, Italy, 2001, September 10-14.

[17] Vennebush, P.B. and Zurkovsky, J., Illuminations - Resources for teaching math, "Sound Sketch Tool", (`http://illuminations.nctm.org/ActivityDetail.aspx?id=36`), accessed 2009/02/29.

[18] Wehinger, R., "Artikulation. Electronic Music. An Aural Score", B. Schott's Söhne, Mainz, Germany, 1970.