Aurally and Visually Enhanced Audio Search with SoundTorch

Sebastian Heise
Hochschule Bremen
Flughafenallee 10
28199 Bremen, Germany
sebastian@h3e.eu

Michael Hlatky
Hochschule Bremen
Flughafenallee 10
28199 Bremen, Germany
michael.hlatky@gmail.com

Jörn Loviscach
Hochschule Bremen
Flughafenallee 10
28199 Bremen, Germany
jl@j3L7h.de

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA
ACM 978-1-60558-247-4/09/04.

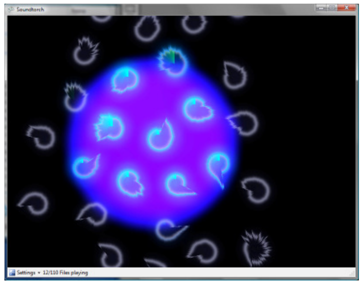


Figure 1: SoundTorch plays back all sounds that are "illuminated" by the virtual flashlight in real time, accompanied by a GPU-based visualization.

Abstract

Finding a specific or an artistically appropriate sound in a vast collection comprising thousands of audio files containing recordings of, say, footsteps, gunshots, and thunderclaps easily becomes a chore. To improve on this, we have developed an enhanced auditory and graphical zoomable user interface that leverages the human brain's capability to single out sounds from a spatial mixture: The user shines a virtual flashlight onto an automatically created 2D arrangement of icons that represent sounds. All sounds within the light cone are played back in parallel through a surround sound system. A GPU-accelerated visualization facilitates identifying the icons on the screen with acoustic items in the dense cloud of sound. Test show that the user can pick the "right" sounds more quickly and/or with more fun than with standard file-by-file auditioning.

Keywords

Sound effects, Foley art, music information retrieval, audio spatialization

ACM Classification Keywords

H5.2. Information interfaces and presentation: User Interfaces: Interaction Styles. H5.5. Information interfaces and presentation: Sound and Music Computing: Systems.

Introduction

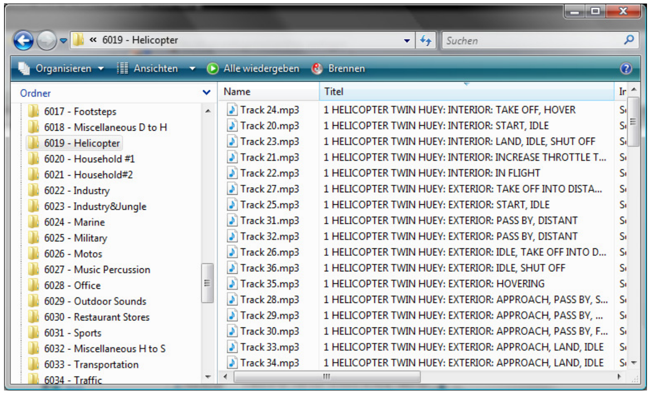
Audio files tend to come in numbers of hundreds and thousands. This does not only concern listeners, but also causes headaches for musicians, sound engineers, and Foley artists (who create effect sounds for movies and radio plays). We address this specific problem: finding effect sounds in large collections such as the freely available www.freesound.org. This task is a daily issue in a number of creative and not-so-creative professions, easily comprehensible to laymen, and can be generalized to other application fields, as discussed in the outlook section of this paper.

We demonstrate a multimodal solution called “SoundTorch” (see Figure 1) that combines methods from music information retrieval (MIR), sonification, and visualization to leverage the perceptual potential of the human brain. This synergetic effect addresses a central topic in research into human-computer interaction: how to provide a way of interaction that elegantly handles the inevitable deviations and outright errors that occur when computer software is used to support search.

SoundTorch has been developed in a university project on applied research. We currently are in talks with several companies about commercial uses. The system as presented in the demo has evolved further from its state reported before [6]. In particular, the icons on the 2D plane formerly only acted as level indicators by flashing. In the version presented here, we generate complex animated shapes that can convey more information and are aesthetically much more pleasing.

Problem Setting

For smaller numbers of audio files, one can get a long way with meaningful file names and tags. The larger a collection gets and the more people contribute to it (both factors are invariably intertwined), the more difficult it becomes to maintain a consistent naming and tagging scheme—if one ever manages to set up such a scheme in the first place. First, it is hard to create a taxonomy that deals with all intricacies of the real world, see Figure 2. Second, sound effects are easily repurposeable, which is heavily employed by Foley artists: Powerline hum can become the sound of a light saber; a mangled package of corn starch produces the sound of footsteps in snow. Such ambiguities are hard to cater for through names and tags. What matters in the end is the sound as such, not how it has been generated.



6017 - Footsteps
6018 - Miscellaneous D to H
6019 - Helicopter
6020 - Household #1
6021 - Household#2
6022 - Industry
6023 - Industry&Jungle
6024 - Marine
6025 - Military
6026 - Motos
6027 - Music Percussion
6028 - Office
6029 - Outdoor Sounds
6030 - Restaurant Stores
6031 - Sports
6032 - Miscellaneous H to S
6033 - Transportation
6034 - Traffic

Name	Titel	Ir
Track 24.mp3	1 HELICOPTER TWIN HUEY: INTERIOR: TAKE OFF, HOVER	S
Track 20.mp3	1 HELICOPTER TWIN HUEY: INTERIOR: START, IDLE	S
Track 23.mp3	1 HELICOPTER TWIN HUEY: INTERIOR: LAND, IDLE, SHUT OFF	S
Track 21.mp3	1 HELICOPTER TWIN HUEY: INTERIOR: INCREASE THROTTLE T...	S
Track 22.mp3	1 HELICOPTER TWIN HUEY: INTERIOR: IN FLIGHT	S
Track 27.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: TAKE OFF INTO DISTA...	S
Track 25.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: START, IDLE	S
Track 31.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: PASS BY, DISTANT	S
Track 32.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: PASS BY, DISTANT	S
Track 26.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: IDLE, TAKE OFF INTO D...	S
Track 36.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: IDLE, SHUT OFF	S
Track 35.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: HOVERING	S
Track 28.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: APPROACH, PASS BY, S...	S
Track 29.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: APPROACH, PASS BY, ...	S
Track 30.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: APPROACH, PASS BY, F...	S
Track 33.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: APPROACH, LAND, IDLE	S
Track 34.mp3	1 HELICOPTER TWIN HUEY: EXTERIOR: APPROACH, LAND, IDLE	S

Figure 2: Standard collections of sound effects require elaborate naming and tagging systems (Screenshot: Sound Ideas “The General” Series 6000).

Music Information Retrieval (MIR) offers methods that can help: Sounds can be classified and searched for by acoustic similarity, partially supported by visualization [3]. However, these methods are clearly limited in that they produce imprecise or wrong results in a substantial number of cases. In particular, the recognition of music genres based on the acoustic content alone is known to hit a “glass ceiling” at about 80 percent recognition rate [4], part of which may be due to semantic information that is not present in the waveforms, such as the artist’s name, or cannot be extracted yet, such as the lyrics. A great deal of uncertainty in MIR results, however, probably still stems from technological issues in feature extraction and pattern recognition.

The limited precision of MIR methods is only one reason why one has to audition dozens of files one by one (and may still miss an interesting candidate). Another reason is that MIR methods need a prototypical audio file to start comparing with—if one does not trust file names or tags to this end. In this case, finding the very first file to compare with requires a manual search.

Overview of the Proposed Solution

The idea behind SoundTorch is to unite methods from MIR with a multimodal user interface that is tolerant concerning misinterpretations by the acoustic analysis and that allows the user to rapidly listen to sounds instead of traversing directory by directory, file by file.

The collection of audio files is depicted as an irregular pattern of animated shapes on the dark background of a 2D display area. The spatial pattern is based on an acoustic analysis that predicts how similar two audio files sound. These data are represented through the spatial distance on the screen. This mapping from the wave form to geometric location employs Mel-Frequency Cepstral Coefficients and a self-organizing map, standard tools from Music Information Retrieval, which typically are applied to visualize collections of music files as landscapes of genres [3].

The user can employ the computer’s mouse or a Nintendo Wii Remote controller to steer an illuminated disk over the screen, imitating a flashlight. All sounds that fall inside this virtual light cone are played back in parallel, without delay and even in high numbers, up to hundreds. The audio transitions when moving the light cone are glitch-free; there is a fluent motion and no hard switching in both graphics and sound.

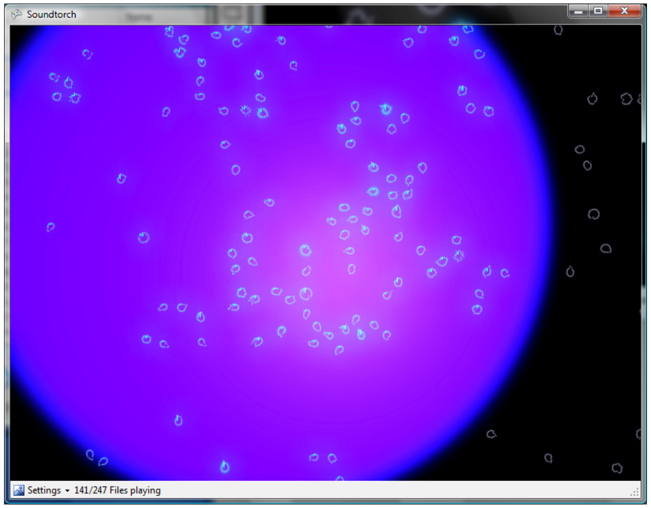


Figure 3: SoundTorch allows zooming out to look at and to audition a collection of thousands of sounds.

The user can easily control the size of the light cone and the zoom factor of the display. A single motion allows changing from a sweeping overview with thousands of files (see Figure 3) to a close examination of a handful of sounds—and back again. Being able to work meaningfully at different levels of detail makes SoundTorch a zoomable interface [1].

The acoustic presentation of several audio files in parallel leverages the potential of the human brain to home in on specific target patterns inside a mixture [2]. This capability surfaces for instance in the cocktail party effect: One can easily focus on the voice of the dialog partner even when dozens of chatting persons are standing in direct vicinity. To fully exploit this capability, the audio output is presented through a surround sound system. The "illuminated" audio files are mapped to the speakers around the user as though one would sit in the center of the light cone on the computer's screen. Sound objects above the current mouse pointer will be played back from the center speaker; objects left and right from the cursor are mapped to the left and right channels, etc. No 3D audio simulation through headphones is employed, as this typically leads to confusion between front and back sounds.

Exploring sound effects works fine because they mostly show little evolution over time. Technically, SoundTorch can also be applied to music files. In music, however, semantics is often more important than the acoustic content. On top of that, music is essentially a blend of many sources, which limits the number of files that can be played simultaneously through acoustic saturation.

SoundTorch provides a connection between vision and sound: The icons representing the audible sounds flash in the rhythm of their acoustic content. In addition, the temporal evolution of the audio power of the signal is mapped around a circle. This creates the icon shapes that mimic bacteria, see Figure 1. As long as a file is played back, its icon rotates so that the current playback location in the file always points upward. This visualization and also parts of the audio computation make heavy use of the graphics processing unit. The visualization allows identifying single sounds even in crowded layouts: As the sounds play asynchronously, the user can easily spot which flashing icon belongs to which sound he or she hears.

Related Work

Up to the authors' best knowledge, the proposed solution for search is currently unique in its combination of advanced auditory and visual output with a fleet-footed zoomable interface. Parts of the solution, however, can be found in systems proposed by other researchers.

Attempts to exploit the human brain's capability for auditory analysis have some history: In 1998, Schmandt proposed an *Audio Hallway* [9]. The user walks along a virtual hallway passing rooms on the left and the right from which he or she hears snippets of news stories. The *Sonic Browser* [5] of Fernström and his coworkers supports listening to several sounds simultaneously. Soundfiles are laid out on a 2D surface; all objects within an "aura" around the mouse pointer are played back. *Sonic Browser* does not employ advanced visualization, zooming, or surround sound; the spatial distribution of the sound objects relies mostly on basic properties of the sound files.

Heuten et al. [7] use a flashlight-based approach to enable visually impaired users to access geographical maps through spatialized acoustic icons such as dabbling water or singing birds. An audio-only interface has been developed by Stewart et al. [11] in 2008. The user can navigate through a virtual 3D space using Nintendo's Wii Remote, pre-listen to several music tracks and select one for full listening. The haptic feedback of the Wiimote tells the user if a music track is near enough for selection.

The simultaneous playback of different audio components may also be applied for creative reasons rather than to exploit perception: *StockSynth* [8] by the Ixi group is intended for real-time performances using a virtual microphone with adjustable scope that is steered with the mouse over the computer screen and “picks up” nearby sound files. The file icons, however, are either laid out in a square grid or have to be defined manually. This system is not intended to produce surround sound. Streich and Ong [12] automatically lay out rhythm loops in 2D according to their similarity. These loops can be played back in parallel by switching them on or off individually to test how well they blend. In CataRT by Schwarz et al. [10] the user selects sound grains that are to be concatenated. This selection involves defining a disk in a 2D feature parameter space.

User Study

We conducted a user test on the original SoundTorch system with less advanced visualization [6]. After familiarizing themselves with the software prototype, 15 subjects completed two different kinds of searches with both SoundTorch and the list-based auditioning function of the audio editing software Sony Vegas 8, where we replaced the file names with random numbers.

First, the subjects had to find a specific sound from different collections of sounds. The target sounds such as a “blib” and a “ding dong” were presented acoustically. SoundTorch turned out to speed up the search by a factor of four. Almost all users actually managed to find the given sounds, where only about half of them succeeded with the list-based interface.

Second, the participants looked for raw material to create imaginary sounds that do not exist as sound files. The examples we used were to find raw material for a dinosaur sound in 100 samples including animal sounds from pigs, sheep, and cows, but also some machine and nature sounds, and to find material for an alarm sound in an extraterrestrial spacecraft from 100 sounds comprising synthesized noises, servo motors, noises from sport events, and war sounds. When using SoundTorch, the subjects reported to enjoy exploring the given corpus of audio files. All subjects compared different sounds with another, trying to find the optimal one; eventually, all told us they were satisfied with the file they found. On average, this took three minutes. In contrast, most users reported to be bored by the list-based auditioning tool. All subjects finished the task when they encountered the first appropriate sound, which took one minute on average; only two thirds of the subjects found the result of their search pleasing.

Conclusion and Outlook

SoundTorch turned out to be an entertaining user interface that speeds up the search for specific sounds and enhances the serendipity in searches for sounds that are appropriate in a given setting. It achieves this by presenting the results of a computerized analysis in a multimodal way that leverages the human brain’s potentials for auditory and visual perception. The extracted and processed data are mapped to sensory stimuli that are easily perceived as well as processed and are tolerant enough to cope with the issue that the computerized analysis never produces accurate results since it inevitably lacks a complete model of the real world including the user’s experiences and preferences.

Standard naming and tagging techniques face the difficulty of acoustic ambiguities that become obvious in the heavy repurposing in Foley art, e.g., corn starch treated to sound like a walk in the snow. These resemble homophones in speech. In contrast, the content-based MIR methods used in SoundTorch face semantic ambiguities, which resemble the synonyms in speech: A collection of door slams may contain some noises that sound like gun shots and others that sound like thunderclaps. This makes it hard to form a mental model of the acoustic landscape of sounds that the SoundTorch interface presents: The different regions could not reliably be named with terms such as "footsteps" or "explosions," but actually would need to carry names that describe the sound as such, not its generation. These names tend to be onomatopoeic such as "clunk" or "splat." We are addressing this in ongoing work.

SoundTorch can be considered a system for combined auditory and visual data mining. We expect that the basic ideas can be carried over to other applications, in particular to search for music—be it canned music for professional film production or be it a home user's collection of MP3 files. This application would benefit from starting the playback of every audio file from a characteristic position such as the chorus or to use "audio thumbnails," i.e., typical snippets instead of the full files. Established methods of MIR can provide such information in a sufficiently robust manner.

Sonification of data may be employed to support search even in domains such as literature or geography. For instance one may form acoustic tag clouds through speech synthesis of salient words or phrases extracted from text documents, in a spirit similar to [9]. Geographic searches may be accompanied by actual or iconified sounds of places, comparable to [7].

References

[1] Bederson, B. and Meyer, J. Implementing a zooming user interface: experience building Pad++. *Softw. Pract. Exper.* 28, 10 (1998), 1101-1135.

[2] Bregman, A.S. *Auditory Scene Analysis.* MIT Press, Cambridge, MA, USA, 1990.

[3] Cooper, M., Foote, J., Pampalk, E. and Tzanetakis, G. Visualization in audio-based music information retrieval. *Comput. Music J.* 30 (2006), 42-62.

[4] Downie, J.S. The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research. *Acoust. Sci. & Tech*. 29, 4 (2008), 247-255.

[5] Fernström, M. Reflections on Sonic Browsing: comments on Fernström and McNamara, ICAD 1998. *ACM Trans. Appl. Perception* 2, 4 (2005), 500-504.

[6] Heise, S., Hlatky, M. and Loviscach, J. SoundTorch: quick browsing in large audio collections. In *Proc. 125th Convention of the Audio Engineering Society* (2008), Paper 7544.

[7] Heuten, W., Henze, N. and Boll, S. Interactive exploration of city maps with auditory torches. In *CHI 2007 Interactivity*, 1959-1964.

[8] Magnusson, T. Screen-based musical interfaces as semiotic machines. In *Proc. NIME 2006*, 162-166.

[9] Schmandt, C. Audio hallway: a virtual acoustic environment for browsing. In *Proc. UIST 1998*, 163-170.

[10] Schwarz, D., Britton, S., Cahen, R. and Goepfer, Th. Musical applications of real-time corpus-based concatenative synthesis. In *Proc. ICMC 2007*, 47-50.

[11] Stewart, R., Levy, M. and Sandler, M. 3D interactive environment for music collection navigation. In *Proc. DAFx-08*, 13-17.

[12] Streich, S. and Ong, B.S. A music loop explorer system. In *Proc. ICMC 2008*.