Searching for Audio by Sketching Mental Images of Sound

A Brave New Idea for Audio Retrieval in Creative Music Production

Peter Knees
Dept. of Computational Perception
Johannes Kepler University Linz, Austria
peter.knees@jku.at

Kristina Andersen
Studio for Electro Instrumental Music (STEIM)
Amsterdam, the Netherlands
kristina@steim.nl

ABSTRACT

We propose a new paradigm for searching for sound by allowing users to graphically sketch their mental representation of sound as query. By conducting interviews with professional music producers and creators, we find that existing, text-based indexing and retrieval methods based on file names and tags to search for sound material in large collections (e.g., sample databases) do not reflect their mental concepts, which are often rooted in the visual domain and hence are far from their actual needs, work practices, and intuition. As a consequence, when creating new music on the basis of existing sounds, the process of finding these sounds is cumbersome and breaks their work flow by being forced to resort to browsing the collection. Prior work on organizing sound repositories aiming at bridging this conceptual gap between sound and vision builds upon psychological findings (often alluding to synaesthetic phenomena) or makes use of ad-hoc, technology-driven mappings. These methods foremost aim at *visualizing* the contents of collections or individual sounds and, again, facilitating *browsing* therein. For the purpose of indexing and querying, such methods have not been applied yet. We argue that the development of a search system that allows for visual queries to audio collections is desired by users and should inform and drive future research in audio retrieval. To explore this notion, we test the idea of a sketch interface with music producers in a semi-structured interview process by making use of a physical non-functional prototype. Based on the outcomes of this study, we propose a conceptual software prototype for visually querying sound repositories using image manipulation metaphors.

CCS Concepts

•**Information systems** → **Speech / audio search;** *Music retrieval;* •**Human-centered computing** → *Graphical user interfaces;*

Keywords

audio retrieval, retrieval by sketch, cross-domain retrieval, music production

ICMR'16, June 06 - 09, 2016, New York, NY, USA

ISBN 978-1-4503-4359-6/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2911996.2912021

1. INTRODUCTION AND CONTEXT

Today's music production is overwhelmingly based on pre-recorded or live-generated sound material. The task of composition consequently often consists of combining sound loops and samples with synthesized and processed elements using a so-called digital audio workstation (DAW), an electronic device or computer application for recording, editing and producing audio files [32]. These types of interfaces typically do not require much prior formal knowledge on music theory or composition (such as traditional tonal harmony), in order to be used. In practice, this development has not only changed the way music is produced on a technical level and democratized music composition but also led to new sound aesthetics and the establishment of electronic music as a dedicated art form and music style, cf. [4, 3]. As results of these changes in practice and mainstream music aesthetics, today, the music products industry has become a major economic force and electronically produced music can be found in all areas of the creative industries (see, e.g., [28] for references).

1.1 The Need for Sound Organization

Working in the context of the EU-funded GiantSteps project [27] we engage with musicians and music producers in order to learn about the processes involved in professional music creation and to improve tools used by practitioners, thereby allowing them a strong peer position in the conceptualization and evaluation of new music interfaces. When asked about their wishes and expectations for future technological developments [1], most of these expert users mention very practical requirements for storing and retrieving sound material and the insuffiency of current solutions for sound search due to the sheer amount of available content:

> "Because we usually have to browse really huge libraries [...] that most of the time are not really well organized." (TOK003)
>
> "If you have, like, a sample library with 500,000 different chords it can take a while to actually find one because there are so many possibilities." (TOK015)
>
> "Like, two hundred gigabytes of [samples]. I try to keep some kind of organization." (TOK006)
>
> "You just click randomly and just scrolling, it takes for ever!" (TOK009)
>
> "I easily get lost... I always have to scroll back and forth and it ruins the flow when you're playing" (PA011)

Figure 1: Screenshots of the Avid Pro Tools 9 [53] (top) and the Ableton Live 8 [52] (bottom) DAWs.

Even from this small selection of statements it becomes clear that organization of audio libraries, "semantic" indexing, and efficient retrieval plays a central role in the practice of music creators and producers. However, with the search tools provided by existing DAWs, this aspect is apparently addressed insuffiently.

1.2 Sound Retrieval in DAWs

The user interfaces of different DAWs are typically very similar. Despite the importance of effective retrieval for users, the access to sample libraries and other audio material is literally marginalized in most DAWs. For composition, the most part of the screen is devoted to the arrangement of different tracks which either consist of sequences of sound segments (displayed as blocks and often additionally overlaid with the waveform of the sound [21]) or piano roll segments representing notes that are synthesized using a virtual instrument (see top of fig. 1). In a live setting, i.e., when an artist is performing on stage, this space is instead often devoted to controlling and mixing the pre-composed tracks (see bottom of fig. 1). In the left and/or right margins of the screen, one can typically find repositories of instruments, sound effects, and sound material (samples), organized in alphabetic lists according to filenames or assigned tags. Additionally, thumbnails of the waveforms might be displayed. In addition to a text search field that matches against the displayed identifiers, semantic tagging of the repositories enables the user to narrow down the search by applying filters. However, in reality, collections remain largely untagged, because tagging needs to be done manually for personal material which is a time consuming and tedious task. Ultimately, to find a suited sample, the filtered list needs to be examined entry by entry by listening to the sounds.

1.3 Proposed New Retrieval Paradigm

In order to organize sounds and make sound collections more intuitively accessible, the challenges of "semantic" description and visualization are central (e.g., [30, 22]). In fact, the visual dimension itself already plays a central role in the verbal description of sound. For instance, sound is characterized through terms such as tone color (timbre), chroma (as an attribute of pitch), or texture, which are all visual metaphors. Such visual metaphors for description of sound were also present during interviews we made:

> "Sometimes I don't have all of my kick drums in my head, because there might be 2,000, or I may not remember what it was but I know I am looking for a soft round sound which is short and dry." (TOK002)

Here, our expert user is referring to textural properties (soft, dry), as well as to properties of shape (round) to describe the qualities of the desired drum sample. Since there is evidence that, when thinking of sound, people (and in particular musicians) might hold a mental image of that sound, we infer that allowing users to *retrieve* a sound by visually sketching it, would allow them to find what they are looking for effectively. Hence, in this paper, we propose a *query-by-visual-sketch* paradigm for sound retrieval.

In order to support and contextualize this idea of utilizing sketches of mental images as queries, in sec. 2, we discuss the requirements and perception of our users and connect these with psychological findings on the correspondence of sound and vision. Within music information retrieval (MIR) research, several related techniques for sound retrieval have been proposed — including acoustic sketches (query-by-example) [25, 41], visualization of sound qualities [30, 8, 22], and map-based arrangements of sounds for browsing [38, 22, 19] — which we discuss in sec. 3.

In sec. 4, we propose a non-functional physical prototype to test the idea of visual querying with users and gather practical feedback. Finally, pointing in the direction of future work, we present a concept for a software prototype which incorporates the desired functionality in sec. 5.

2. MENTAL IMAGES OF SOUND

One of our goals is to understand our participants' mental models of sound and music. If we understand the musical perceptions of our users, we may in turn be able to suggest new ways to find sound. To this end, we explore the mental models of sounds of our users by asking them to illustrate the sound they would like to be able to make by drawing it with coal on a small piece of paper.[1] Each participant spends time explaining the illustrations, and this opens up a conversation about their mental model of sound:

> "I don't wanna see the wave, I want to use my ears and as soon as I start being able to see the wave I start getting into this visual world where the whole, my brain totally changes and the way I interact with, what I am doing totally changes." (STE023)

In order to further inspire lateral thinking, some participants are asked to make this drawing on the palm of their

[1] cf. http://ears2.dmu.ac.uk/learning-object/drawing-sounds

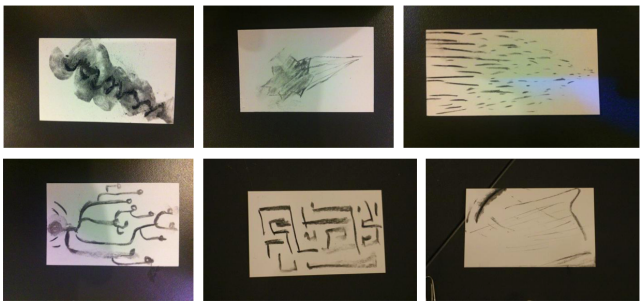

Figure 2: Coal drawings of mental images of sound.

own hand with a black marker. This has the advantage of making the task both much more difficult, but — maybe counter intuitively — also easier, in the same way that an elaborate surrealist art game such as Exquisite Corpse allows easy expression and complex outcomes [5].

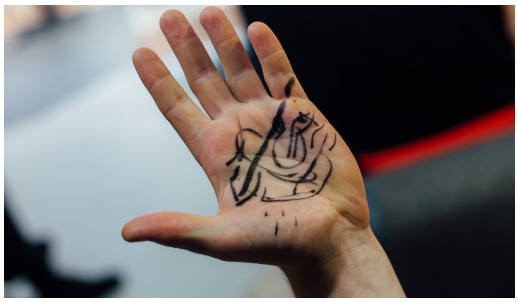

Figure 3: Drawing of "the sound you want to be able to make" on the hand.

> "So it would be really useful to have some kind of sorting system for drums, for example, where I could choose: 'bass drum', and here it is 'bass' and 'bright', and I would like it to have, maybe, bass drum 'round' and 'dry', and you can choose both..." (TOK002)
>
> "It would be even more useful to be able to search for a particular snare, but I can't really imagine how, I need something short, low in pitch, dark or bright in tone and then it finds it..." (TOK003)
>
> "There are a lot of adjectives for sound, but for me, if you want a bright sound, for example, it actually means a sound with a lot of treble, if you say you want a warm sound, you put a round bass, well, round is another adjective." (TOK009)

In these quotes, sounds are described by shapes (round), brightness (bright, dark) and textures (soft, dry). While these might be regarded as unusual descriptors of sound, there is some evidence that many humans make, to some degree, use of synaesthetic connections between visual perceptions and sound.

When talking about "synaesthesia" in this artistic context, it needs to be clarified that we use it as "an aesthetic appropriation of the neurological condition in which stimulation of one sensory modality triggers involuntary sensation in another" [11], which is a rare, asymmetric, and individual phenomenon. Thus, we make use of a weak definition of synaestesia as cross-modal associations, cf. [17, 22], and, in the context of computer science, "the more general fact that digital technologies offer, if not a union of the senses, then something akin: the intertranslatability of media, the ability to render sound as image, and vice versa." [11]

Some experiments have found an association of *brightness* with musical scales, modes, and pitch height [10]. Datteri and Howard find an association of the frequency of pure sine tones and color frequency [12], while Rusconi et al. point out that some people associate pitch height with spatial height [42]. As mentioned by Datteri and Howard [12], Marks finds that some participants associate an increase of brightness of grey surfaces with increase in loudness of pure tones, while others associate an increased loudness with a decrease of brightness, and suggests that most participants associate visual brightness with auditory brightness [35]. This diversity is an indication that a universal function to translate visual impressions into sound that generally matches human perception may not exist. In terms of *shape*, there is repeatable evidence of object-sound associations, For example, it was shown that nonsense words such as *baluba*, *maluma*, or *bouba* are associated with rounded, while words such as *takete* or *kiki* are associated with spiky and angular shapes [29, 36].

We find an interesting opportunity in the relation between our users' mental images of sound and the possibilities for using these intuitions for sound retrieval. The emerging goal of this process is to make use of these connections for building a visual-query sound search engine. In the following, we review how this area has been discussed in related work in multimedia retrieval.

3. RELATED WORK: VISUAL — SOUND — RETRIEVAL

Three areas are touched by the idea of using graphical sketches of mental images of sound for audio search: vision, sound, and retrieval. Extensive work — both academically and artistically — has been conducted in all these areas. To discuss the most relevant examples of this related work, in the following, we focus on work at the intersections, precisely, works at the intersection of visual and sound *(VS)*, visual and retrieval *(VR)*, sound and retrieval *(SR)*, and the intersection of all three *(VSR)*, cf. fig. 4.

3.1 Visualization/Sonification

At the intersection of visual and sound, we find work that visualizes sound or sonifies images, or both.

The area of **music visualizations** is a wide field and ranges from software and plug-ins for automatic visualizations of played back music (e.g., the iTunes Visualizer), to stage show effects, to visuals in music performance, to the dedicated VJ culture. The artistic discussion and discourse at the intersection of visual arts and music composition and the motif of synaesthesia, historically, has influenced and been influenced by the works of Kandinsky, Schönberg, Cage, Stockhausen, or Xenakis, cf. [20], to name a few prominent exponents, and is a reoccurring theme of art exhibitions, e.g., [14, 40, 37]. The abstract musical animations by Oskar Fischinger have even reached the mainstream entertainment industry (e.g., through a visualization of Bach's Toccata and Fugue in Walt Disney's Fantasia) [26].

In the field of computer science, various strategies for vi-

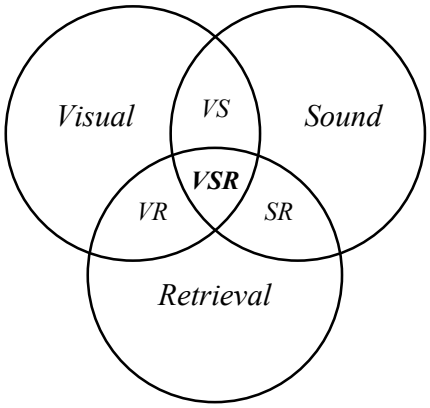

Figure 4: Venn diagram of areas of related work. Our proposed approach to audio retrieval through visual sketches belongs to the central area (*VSR*).

sualizing sounds have been proposed. *ThumbnailDJ* [8] produces visual summaries for audio files using repurposed music notation symbols to display information on aggressiveness, bass, tempo, volume, and genre for DJs. Genre information is also mapped to colors for fast distinction. *Music Icon* [30] generates expressive file system icons for music files using blooms with two rings of petals that reflect music features by color, shape, and number of petals. In order to map musical properties to the eight parameters used to generate a bloom, a neural network is trained using a small set of hand-crafted examples. Grill et al. [22, 23] visually model perceptual qualities of textural sound. In the visualization, they combine two aspects, namely mapping the overall structure of a sound collection and visualizing the qualities of its elements. Since they propose a user interface for visual browsing, we discuss this work in more detail in sec. 3.4.

In the area of **image sonification**, i.e., the generation of non-speech sound or music that conveys information of visual data (cf. [31] for a general definition), the process of visualization is inverted. Existing work is dealing with extraction of visual features, e.g., color, brightness, texture, or shape, and using these to control acoustic properties. Thoret et al. [49], for instance, sonify two-dimensional curves by imitating the sounds of friction of a pencil drawing the shape to sonify. Other approaches often aim at automatically composing music from images, interpreting the two dimensions of images onto two acoustic dimensions, e.g., the position of pixels on the y-axis as pitch and the position along the x-axis as point in time. Lauri Gröhn utilizes a cell-automaton-like concept to filter images by removing pixels in an iterative process.[2] *Bondage* by Atau Tanaka is an interactive installation including the transformation of a displayed picture into sounds.[3] The *Monalisa* project [24] transforms sound into image data and vice versa using software plugins that allow the approach to apply sound effects to image data, and to apply image effects to sounds. *sound/tracks* [39] generates a musical composition from the images of a passing scenery (recorded from trains) by mapping spatial height to octaves and colors to notes using the synaesthetic mapping of composer Alexander Skrjabin's color keyboard.

[2]http://www.synestesia.fi

[3]http://www.ataut.net/site/Bondage

3.2 Image Search and Sketch Retrieval

Content-based retrieval in the visual domain is a well-researched area, e.g., [33]. In the context of this paper we are foremost interested in methods for sketch-based and shape-based retrieval [51]. In sketch-based image retrieval, the goal is to use a schematic (even binary) sketch as query to find similar full color images. Existing methods build upon local gradient histograms, structure tensors [15], bag of features descriptors [16], or keyshapes [43] and can further be applied for 3D-object retrieval [44] and image synthesis and montage applications [7]. In our proposed approach, extracting features and relevant parameters from sketches and matching them to repositories of richer and more complex data is essential in order to achieve our goal.

3.3 Audio Search and Semantic Indexing

Current audio and music search is dominated by the query-by-example paradigm [6]. An early example of a content-based sound retrieval system is *SoundFisher* [54], which lets the user query a database by combining examples with tags and other meta-data. Another early example uses audio features for indexing, classification, and retrieval of audio/video segments in production studios [55]. Query-by-example is also practiced in *mosaicing*, an effect that "reconstructs" a target sound by concatenating similar sounding slices of other recordings [56, 45, 34]. An alternative to direct query-by-example (audio-to-audio) is query-by-semantic-example, which first aims at deriving semantic tags through classification (see also [47]) and then compares sounds on the level of semantic categories [2].

A coarser approach to querying by example is to use an "acoustic sketch," so to say, i.e., articulating (through a microphone) the qualities of the desired sound with voice. So far, this has been applied for retrieving samples or tracks with a similar beat structure for DJs ("query-by-beat-boxing" [25]) and for sound synthesis [41]. In the context of sound synthesis and creation of textures, also sound-to-(haptic)-image, image-to-texture, and haptics-to-texture transformations have been proposed [13]. However, existing methods are either unimodal, i.e., do not take advantage of synaesthetic correspondence with the visual domain, or focus on the translation of parameters for sound synthesis rather than indexing and retrieval. To the best of our knowledge, there is no work that addresses sound search by graphical querying using synaesthetic information.

3.4 Sound Browsing through Collection Visualizations

Finally, we want to discuss existing methods at the intersection of visual, sound, and retrieval, i.e., methods that assist in retrieving sounds from collections by providing a visual browsing tool. Most proposed sound browsing systems are based on maps, i.e., 2D arrangements of sounds. The *Sonic Browser* [18] gives the user the possibility to assign visual properties such as color, shape, size, and location to characteristics of audio files and then navigate through the resulting visualization plane by simultaneously playing sounds within a selected region through stereo spatialization. Pampalk et al. [38] present a hierarchical user interface to drum sample repositories based on self-organizing maps and

psychoacoustically motivated descriptors of drum sounds. Coleman [9] extracts event-synchronous audio samples from existing music collections and proposes an interactive scatter plot for browsing. Schwarz and Schnell [46] apply a similarity function developed for mosaicing together with a spring model for layout and multi-dimensional scaling to enable the user to interact with and filter the contents of sound effect and sample databases. Fried et al. [19] present *AudioQuilt*, which optimizes 2D arrangements of audio samples based on user preferences and can be used for navigation in snare drum and synth sound databases.

Grill and Flexer [22] visualize perceptual qualities of sound textures through symbols that are arranged on a map to reflect the structure of the collection (see fig. 5).[4] To this end, they map bipolar qualities of sound that describe spectral and temporal aspects of sound, to visual properties. The spectral qualities of pitch (high vs. low) and tonality (tonal vs. noisy) are mapped to brightness/hue and saturation, respectively. The temporal (or structural) qualities of smoothness vs. coarseness, order vs. chaos, and homogeneity vs. heterogeneity are associated with the jaggedness of an element's outline, the regularity of elements on the grid, and a variation in color parameters, respectively. Evaluation of these correspondences through a survey showed that subjects were able to meaningfully associate sounds with the chosen graphical representations. These parameters could also be used in our proposed search system, i.e., to match user-drawn shapes to sounds.

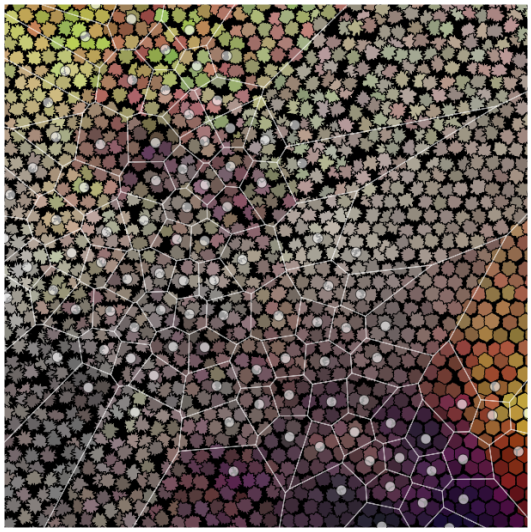

Figure 5: Screenshot of the texture browser by Grill and Flexer [22].

4. A NON-FUNCTIONAL PROTOTYPE TO TEST THE IDEA

In order to explore how our users imagine sound, we decided to build a non-functional prototype. This was the simplest of objects, a cardboard box the size of a hand, a traditional size for a music device or tablet. The purpose of the prototype was to give our idea a physical representation and allow it to function as a conversational prop during interviews. During the interviews, the users were told that this prop would find the sounds they want if they express them visually and asked to do various small exercises: *draw your visualization of a piece of music*, *use small colored pieces of paper to illustrate the piece again*, and *imagine how the box could be used in your particular practice*. This is accompanied by an ongoing conversation about images and sound.

The following quotes are taken from this process, conducted with 21 users at professional music events in Amsterdam and Paris. In the interest of brevity, we will show examples from one user conversation (ST002) here and then summarize the ideas and comments from all the interviews.

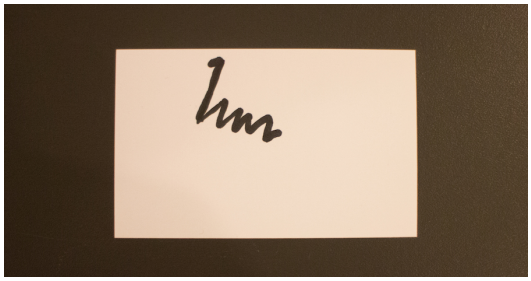

Figure 6: Drawing the sound.

User ST002 describing the sound and fig. 6.

> "For me, it's sort of like a chair moving back and forth. You have the note which is the long line and the *(makes high-pitched noise)*. I think they're actually both at the same time. You sort of perceive the note first and then the *(makes noise)*."

Figure 7: Collaging the sound.

Describing the collage in fig. 7

> "This is how you should view it. The base of the track is like the beat, it's not completely rigid to the grid. It doesn't have a lot of swing so it's divided into four but it has a little bit of space. It's also because there's not a lot of variations, there might be some little fills and everything but the basis is still the same. So you have the basis. Over there, there are some sounds and they're not really... they're pretty sort of neutral, so I tried to stay more straight into the neutral colors like grey and blue and a little bit of red, because sometimes it's a bit more calmed down. They're overlapping, but it's not really totally chaotic. You can, when you look at it, it's like fourth,

[4] http://grrrr.org/test/texvis/map.html

Figure 8: Imagining using the search box.

there's actually a lot going on, but it doesn't actually feel like it. That's what I tried to recreate."

We then start talking about how to use the box, see fig. 8.

> *User ST002*: "Okay, 'I really want a squiggly sound here,' and you get a really round sound and you're like, 'Hey this could also work'. Or you could start modifying your squiggle."
>
> *Interviewer*: "If you could imagine you could constantly change what would come back if it was a live query."
>
> *User ST002*: "Yeah, it could be like, 'I want this sound, okay well this is not what I had in mind, let's try this and this,' then you can really shape your sound to what you want to."

At the end of the interview we talk more generally:

> *Interviewer*: "We had an artist suggest that he was associating particular colors to each of his samples as he made them. He would go 'this is a green sound' but he would have a really detailed opinion about what kind of green."
>
> *User ST002*: "Yeah, no, a lot of artists have that actually. It's really funny. Imagine if their whole song is sort of structures and shapes with colors and everything else."

Over the course of the interviews a number of ideas and notions came up in addition to the importance of brightness, shape, and texture in mental models identified in sec. 2:

- Our expert users have complex mental models of their own sounds and color plays a central role: "I see the music sometimes as more aesthetic and something that I can see more than something that I can hear" (PA013), "When I listen to music I see colors. [...] I remember colors." (PA011), "Different sounds to me have specific colors" (PA009)
- The temporal dimension of sound is important: "a very useful and beautiful way of displaying the track over time, the colors based on timbre and stuff" (ST001)
- For some users, current forms of visualization, especially waveforms, are useful, while others are very dissatisfied: "Listen [...] with the waveform, [is] really useful for me, I can go to the drop directly cause know from the waveform where it is and I can listen to it and then I can jump, like, to the middle section." (TOK011), "The waveform is a beautiful thing to look at, depending on its sound." (TOK014) vs. "You're confronted with these boring waveforms... they bore your eyes... you're already not listening to something because you just see a square wave." (PA008), "If you think about it, most waveforms of songs look like bricks because they've been smashed to shit with brick wall limiters and compression." (TOK007), "Looking at the waveform is not really helpful." (TOK003)
- Users are open to systems that might not work perfectly out of the box but learn their associations over time (personalization): "You could imagine that your computer gets used to you, it learns what you mean by grainy, because it could be different from what that guy means by grainy" (PA008)
- Users would use images as references to find files: "I don't have the actual ability to use images [now], so I just use color." (PA009)

5. PROPOSAL FOR A FUTURE SOUND RETRIEVAL SYSTEM

Based on the findings from our interviews and from surveying the literature, we propose a software interface for sound search based on queries consisting of sketches of mental images. A central requirement for such an interface is that it needs to be able to deal with different sound properties and different types of sounds, such as effects, samples, ambient, tonal, or textured recordings, and therefore needs to comprise *different simultaneous representational models for indexing*. For instance, while tonal aspects might be best represented using symbolic music notation, noise sounds should be modeled primarily via their textural properties. It is expected that modeling and indexing will heavily draw from audio content processing and analysis methods. Methods for source separation will play a particularly important role in order to isolate individual sound objects, cf. [48].

As pointed out before, while some audio-visual associations are more universal (e.g., acoustic representations of shape seem to be more generally agreed on), others, like the correspondence between spatial height and pitch height or the brightness of sound, are more subjective. The association of colors with sound, although appearing to be the most frequent association, seems also to be the most subjective. Thus, another requirement for the sketch search interface is *adaptiveness to the preferences of the user* over time.

A query to the search engine is constructed visually aiming at reflecting the user's mental model of sound (i.e., the information need). To this end, the search interface will resemble more an image manipulation program than a text search engine. Fig. 9 shows a UI mockup of the envisioned system. The central element is the canvas, holding the current query. The user can start with an empty canvas and build the query from scratch or input an existing sound, which is subsequently analyzed, deconstructed, and displayed using the available symbolic representations. While the user is drawing or moving objects on the canvas, the list of search results gets constantly updated, encouraging exploration and modification. The canvas itself features a two-dimensional layout, with the x-axis representing time (or temporal order), and consists of multiple layers, which can be turned on and off selectively, each representing a different type of

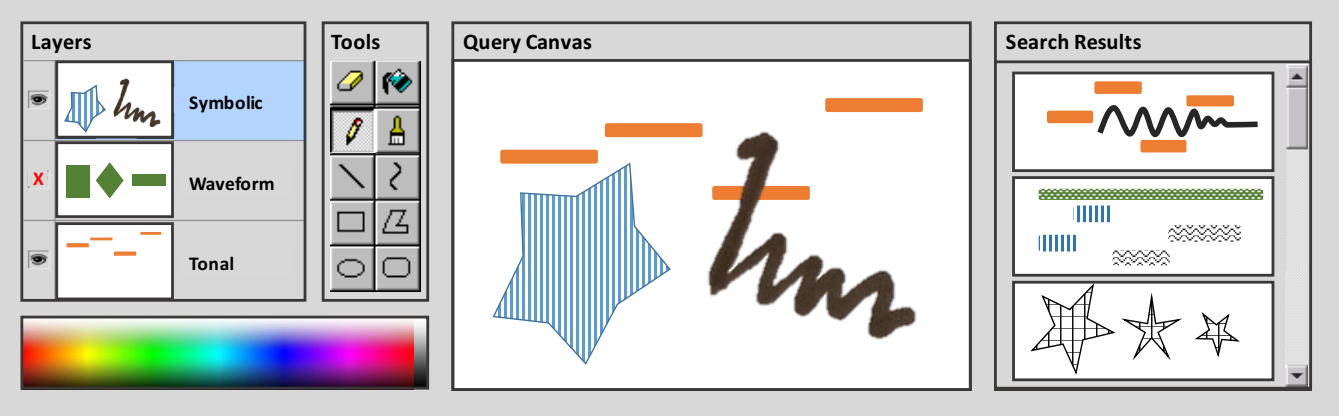

Figure 9: UI mockup of the proposed query-by-visual-sketch search engine for sound.

indexing modality. In the different layers, the x- and y-axes are interpreted differently:

1. *symbolic layer*: sequential order vs. spatial area of symbols
2. *waveform layer*: time vs. amplitude
3. *tonal layer* (or spectral or piano roll layer): time vs. pitch (frequency activation)

For indexing and query matching, i.e., similarity calculation, the symbolic layer requires a segmentation of the audio into sound objects which are individually modelled and then can be expressed as a sequence of symbols (i.e., a string). The mappings used by methods discussed in related work (sec. 3) are promising candidates to create such symbols. However, not all of these mappings are isomorphic in the sense that their expressivity can be inverted and not all relevant acoustic dimensions are covered. The waveform and tonal layers basically hold sketches of the expected waveform and the piano roll (or spectrogram) of the sound, repectively, and can be matched by stretch- and (partly) location-invariant image comparison. Alternatively, the tonal layer can be matched using symbolic music retrieval methods such as melody contour matching, e.g., [50].

For sketching the appearance of the individual layers, common image manipulation tools are provided:

- Draw dots, lines, and paths (e.g., waves, zig-zag)
- Draw shapes: circles, ellipses, rectangles, polygons (e.g., stars)
- Fill shapes with texture (from a predefined set)
- Choose color to draw to represent different sound qualities (e.g., tonality, instrument, mood, or any other semantic label)
- Eraser (selectively remove shapes or parts of drawings)

An important addition to the system is that — in contrast to text- or image-based search — in case retrieval delivers no or only unsatisfactory results, sound synthesis can be applied instead, based on the provided parameters, using, e.g., the methods discussed in sec. 3.3. That is, if the desired sound is not present in the database, it can still be built/synthesized given the visual description.

6. CONCLUSIONS

We proposed a new search paradigm for sound retrieval based on visual sketch queries to replace or complement existing text- and tag-based search engines as exemplified in current digital audio workstations. From a substantial amount of interviews with professional music producers, we identified a need for improved organisation of and access to large sound repositories as well as a common conception of sound through visual metaphors and synaesthetic sensations. We argued for making use of the visual mental models of sound for retrieval and presented a mockup UI to allow for visual sketches using image manipulation metaphors.

While an actual implementation of such a search engine is beyond the scope of this paper, we believe that setting such a target points research in content-based audio retrieval into a highly relevant direction (as evidenced not at least by the explicit demand for this type of functionality by users). Existing work that already builds upon audio-visual correspondences for visualization purposes should be tapped for also facilitating the opposite information flow and giving users graphical tools to "build" the sound they want to find. Audio and music information retrieval research has achieved significant improvements in the last years wrt. sound description and indexing. Therefore, although it is without question a very difficult, subjective, and ambiguous task that requires extensive efforts, we believe that visual sound retrieval does not end with browsing interfaces, but should be taken to the level of search engines, to help people find the sounds they already have in mind.

7. ACKNOWLEDGMENTS

This work is supported by the European Union's seventh Framework Programme FP7 / 2007–2013 for research, technological development and demonstration under grant agreement no. 610591 (GiantSteps).

8. REFERENCES

[1] K. Andersen, F. Grote. GiantSteps: Semi-structured conversations with musicians. In *Proc CHI EA*, 2015.

[2] L. Barrington, A. Chan, D. Turnbull, G. Lanckriet. Audio information retrieval using semantic similarity. In *Proc ICASSP*, 2007.

[3] T. Brett. DJ culture in the mix: Power, technology, and social change in electronic dance music. *Popular Music and Society*, 38(3):389–391, 2015.

[4] B. Brewster, F. Broughton. *Last night a DJ saved my life: The history of the disc jockey.* Grove/Atlantic, 2007.

[5] A. Brotchie, M. Gooding, eds. *A book of surrealist games.* Shambhala Redstone Editions, 1995.

[6] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proc IEEE*, 96:668–696, April 2008.

[7] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, S.-M. Hu. Sketch2photo: Internet image montage. *ACM TOG*, 28(5):124, 2009.

[8] Y.-X. Chen, R. Klüber. ThumbnailDJ: Visual thumbnails of music content. In *Proc ISMIR*, 2010.

[9] G. Coleman. Mused: Navigating the personal sample library. In *Proc ICME*, 2007.

[10] W. G. Collier, T. L. Hubbard. Musical scales and brightness evaluations: Effects of pitch, direction, and scale mode. *Musicae Scientiae*, 8:151–173, 2004.

[11] C. Cox. Lost in translation: Sound in the discourse of synaesthesia. *Artforum Int.*, 44(2):236–241, 2005.

[12] D. L. Datteri, J. N. Howard. The sound of color. In *Proc ICMPC*, 2004.

[13] A. Del Piccolo, S. Delle Monache, D. Rocchesso, S. Papetti, D. A. Mauro. To "sketch-a-scratch". In *Proc SMC*, 2015.

[14] S. Duplaix, M. Lista, B. Veret, C. Delineau, eds. *Sons & Lumières – Une histoire du son dans l'art du XX^e siècle.* Éditions du Centre Pompidou, 2004.

[15] M. Eitz, K. Hildebrand, T. Boubekeur, M. Alexa. A descriptor for large scale image retrieval based on sketched feature lines. In *Proc SBIM*, 2009.

[16] M. Eitz, K. Hildebrand, T. Boubekeur, M. Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE TVCG*, 17(11):1624–1636, 2011.

[17] K. K. Evans, A. Treisman. Natural cross-modal mappings between visual and auditory features. *JOV*, 10(1):6.1–12, 2010.

[18] M. Fernström, E. Brazil. Sonic browsing: An auditory tool for multimedia asset management. In *Proc ICAD*, 2001.

[19] O. Fried, Z. Jin, R. Oda, A. Finkelstein. AudioQuilt: 2D arrangements of audio samples using metric learning and kernelized sorting. In *Proc NIME*, 2014.

[20] K. Giannakis. A comparative evaluation of auditory-visual mappings for sound visualisation. *OSO*, 11:297–307, Dec. 2006.

[21] K. Gohlke, M. Hlatky, S. Heise, D. Black, J. Loviscach. Track displays in DAW software: Beyond waveform views. In *AES 128*, 2010.

[22] T. Grill, A. Flexer. Visualization of perceptual qualities in textural sounds. In *Proc ICMC*, 2012.

[23] T. Grill, A. Flexer, S. Cunningham. Identification of perceptual qualities in textural sounds using the repertory grid method. In *Proc Audio Mostly*, 2011.

[24] K. Jo, N. Nagano. Monalisa: "See the sound, hear the image". In *Proc NIME*, 2008.

[25] A. Kapur, M. Benning, G. Tzanetakis. Query-by-beat-boxing: Music retrieval for the DJ. In *Proc ISMIR*, 2004.

[26] C. Keefer, J. Guldemond, eds. *Oskar Fischinger 1900-1967: Experiments in cinematic abstraction.* EYE Film Instituut, 2013.

[27] P. Knees, K. Andersen, S. Jordà, M. Hlatky, A Bucci, W. Gaebele, R. Kaurson. The GiantSteps project: A second-year intermediate report. In *Proc ICMC*, 2016.

[28] P. Knees, Á. Faraldo, P. Herrera, R. Vogl, S. Böck, F. Hörschläger, M. Le Goff. Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In *Proc ISMIR*, 2015.

[29] W. Köhler. *Gestalt psychology.* Liveright, 1929.

[30] P. Kolhoff, J. Preuss, J. Loviscach. Music icons: Procedural glyphs for audio files. In *Proc SIBGRAPI*, 2006.

[31] G. Kramer, ed. *Auditory display: Sonification, audification, and auditory interfaces.* Addison-Wesley, 1994.

[32] L. Landy. *Making music with sounds.* Routledge, 2012.

[33] M. S. Lew, N. Sebe, C. Djeraba, R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM TOMCCAP*, 2(1):1–19, Feb. 2006.

[34] E. Maestre, R. Ramírez, S. Kersten, X. Serra. Expressive concatenative synthesis by reusing samples from real performance recordings. *CMJ*, 33(4):23–42, Dec. 2009.

[35] L. E. Marks. On associations of light and sound: The mediation of brightness, pitch, and loudness. *AJP*, 87 (1-2):173–188, 1974.

[36] D. Maurer, T. Pathman, C. J. Mondloch. The shape of boubas: Sound–shape correspondences in toddlers and adults. *Dev. Sci*, 9:3:316-322, 2006.

[37] C. Meyer, ed. *Kandinsky, Kupka, Schönberg – Abstraction and atonality.* Museum Kampa, 2011.

[38] E. Pampalk, P. Hlavac, P. Herrera. Hierarchical organization and visualization of drum sample libraries. In *Proc DAFx*, 2004.

[39] T. Pohle, P. Knees, G. Widmer. sound/tracks: Real-time synaesthetic sonification and visualisation of passing landscapes. In *Proc ACM Multimedia*, 2008.

[40] C. Rainer, S. Rollig, D. Daniels, M. Amme, eds. *See this Sound.* W. König, 2009.

[41] D. Rocchesso, G. Lemaitre, P. Susini, S. Ternström, P. Boussard. Sketching sound with voice and gesture. *IX*, 22(1):38–41, Jan. 2015.

[42] E. Rusconia, B. Kwana, B. L. Giordano, C. Umiltà, B. Butterworth. Spatial representation of pitch height: the SMARC effect. *Cognition*, 99(2):113–129, 2006.

[43] J. M. Saavedra, B. Bustos. Sketch-based image retrieval using keyshapes. *MTAP*, 73(3):2033–2062, 2013.

[44] J. M. Saavedra, B. Bustos, M. Scherer, T. Schreck. Stela: Sketch-based 3D model retrieval using a structure-based local approach. In *Proc ICMR*, 2011.

[45] D. Schwarz. Current research in concatenative sound synthesis. In *Proc ICMC*, 2005.

[46] D. Schwarz, N. Schnell. Sound search by content-based navigation in large databases. In *Proc SMC*, 2009.

[47] M. Slaney. Mixtures of probability experts for audio retrieval and indexing. In *Proc ICME*, 2002.

[48] P. Smaragdis. User guided audio selection from complex sound mixtures. In *Proc UIST*, 2009.

[49] E. Thoret, M. Aramaki, R. Kronland-Martinet, J.-L. Velay, S. Ystad. From shape to sound: Sonification of two dimensional curves by reenaction of biological movements. In *Proc CMMR*, 2012.

[50] R. Typke. *Music retrieval based on melodic similarity.* PhD thesis, Univ Utrecht, 2007.

[51] R. C. Veltkamp. Shape matching: Similarity measures and algorithms. In *Proc SMI*, 2001.

[52] Wikipedia. Ableton, 2016. [Online; acc. 02/11/16].

[53] Wikipedia. Pro Tools, 2016. [Online; acc. 02/11/16].

[54] E. Wold, T. Blum, D. Keislar, J. Wheaten. Content-based classification, search, and retrieval of audio. *IEEE MultiMedia*, 3(3):27–36, 1996.

[55] T. Zhang, C.-C. Kuo. Classification and retrieval of sound effects in audiovisual data management. In *Proc Asilomar SSC*, volume 1, 1999.

[56] A. Zils, F. Pachet. Musical mosaicing. In *Proc DAFX*, 2001.