Personal and Ubiquitous Computing (2021) 25:723–736
https://doi.org/10.1007/s00779-020-01388-1

ORIGINAL ARTICLE

A proposal and evaluation of new timbre visualization methods for audio sample browsers

Etienne Richan[1] · Jean Rouat[1]

Received: 9 December 2019 / Accepted: 5 March 2020 / Published online: 19 April 2020

Abstract

Searching through vast libraries of sound samples can be a daunting and time-consuming task. Modern audio sample browsers use mappings between acoustic properties and visual attributes to visually differentiate displayed items. There are few studies focused on how well these mappings help users search for a specific sample. We propose new methods for generating textural labels and positioning samples based on perceptual representations of timbre. We perform a series of studies to evaluate the benefits of using shape, color, or texture as labels in a known-item search task. We describe the motivation and implementation of the study, and present an in-depth analysis of results. We find that shape significantly improves task performance, while color and texture have little effect. We also compare results between in-person and online participants and propose research directions for further studies.

Keywords Timbre perception · Texture synthesis · Color · Shape · Known-item search · Media browsers

1 Introduction

Modern sample libraries can contain thousands of synthesized or recorded sound samples. A common approach when searching for a sample is to filter the contents of the library based on keywords or categories and then audition the resulting samples one by one.

Several media browsers have been developed to accelerate this process by placing samples produced by query results in a scatterplot visualization, or a *starfield display* [4]. Some implementations allow users to specify what metadata or audio descriptors to use as the axes of the display [9, 15], while others use dimensionality reduction (DR) methods to project a high-dimensional set of auditory features to a 2D space [16, 18, 24]. The latter approach provides less meaningful axes but can produce an effective clustering of similar sounds. This feature-based generation of sample coordinates is often augmented by a visual labeling method. These visual labels can help the user to recognize types of sounds or sounds that they have already auditioned. In existing sample browsers, colors are mapped to timbre features [9, 16–18], and shapes are used to either distinguish categorical variables (e.g., instrument type) [9] or to visualize time-varying features [18, 25].

1.1 A novel sample browser with textural labels

We developed a sound sample browser which lets users visually label sounds using textural images. It uses a pretrained neural network model [34] for artistic style transfer [19] to synthesize visual labels for all samples in the library based on a reference set of sound-image pairs. Users can simply choose textural images that they wish to associate with certain sounds and the software takes care of the rest. The advantage of using this method is that no explicit mapping between sound descriptors and visual parameters needs to be defined. We think this is an interesting alternative to the more common approach of associating specific audio features to shape or color parameters. Our browser also uses dimensionality reduction of timbre features to place samples in the interface.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00779-020-01388-1) contains supplementary material, which is available to authorized users.

✉ Etienne Richan
etienne.richan@usherbrooke.ca

Jean Rouat
jean.rouat@usherbrooke.ca

[1] NECOTIS GEGI, CIRMMT, Université de Sherbrooke, Sherbrooke, Canada

1.2 Research questions

While we wish to evaluate the design choices of our sample browser, to what degree *any* of these types of visual labels improve sound search is still an open question. We designed our study to address the following questions:

- Is there a difference between using color, shape, or texture as visual labels in an audio sample browser?
- Does our timbre feature–based placement method assist search, and if so, are these visual labels effective when the information provided by placement is removed?

This article presents the design of the study as well as methods for generating visual labels and placing musical samples based on their timbre features. We provide in-depth analysis of results from a group of 15 participants who were recruited to perform the study in a controlled environment. We find that the placement method and shape labels improve participant efficiency, but do not significantly improve task completion time. We also compare these results with a second group of 14 participants who performed the study remotely, but find few commonalities between the two groups.

2 Related work

The field of information visualization provides a rich resource of theory and guidelines for visual label design. Borgo et al. [7] provide an extensive review of glyph visualization, a general form of label designed to communicate values visually. Chen and Floridi [13] propose a taxonomy for four types of visual channels: geometric channels (e.g., size, shape, orientation), optical channels (e.g., hue, saturation, brightness, texture, blur, motion), topological and relational channels (e.g., position, distance, connection, intersection), and semantic channels (e.g., numbers, text, symbols).

While visualization theory principles can be applied to arbitrary sources of information, visualizations of sound can benefit from visual metaphors that appeal to intuitive associations we might make between acoustic and visual properties.

2.1 Cross-modal correspondences for timbre visualization

Studies of cross-modal correspondences provide useful insights into audio interface design as they highlight associations between vision and audition that a large part of the population might intuitively understand. The *kiki-bouba* experiment [59] is an early study of such correspondences which found that across cultures and ages, most people associate the vocalized word *“bouba”* with rounded shapes and *“kiki”* with pointed shapes. Recently, an investigation of the cross-modal correspondence of timbre and shapes [2] came to a similar conclusion with regard to musical timbre.“Soft” timbres were associated with round shapes, while “harsh”, brighter timbres were associated with spiky shapes. The same work also highlighted a tendency to associate the soft timbres with blues and greens and harsh timbres with reds and yellows.

Giannakis and Smith [20] studied correspondences between acoustic descriptors and visual parameters, furthering work begun by Walker [54] on associations between pitch, loudness, and visual features such as size, position, and lightness. With *Sound Mosaics*, they studied associations between synthesized timbres and textural images containing repeated elements with varying parameters such as coarseness, distribution, and granularity. They found strong associations between granularity and spectral compactness as well as between coarseness and spectral brightness.

Grill and colleagues performed a study highlighting several high-level perceptual qualities of textural sounds [23] and proposed visualizations [22] as well as methods for extracting descriptors [21] for each one. Two of their proposed perceptual metrics, height and tonality (measuring whether a sound is more tone-like or noise-like), are quite similar to the timbral descriptors for brightness and spectral flatness. Both were visualized using color: height was mapped to a range of hue and brightness ranging from bright yellow (high) to dark red (low) and saturation was mapped to tonality.

Berthaut et al. [6] as well as Soraghan [51] studied potential correlations between acoustic properties and those of animated 3D objects. The former found a preference for associating the spectral centroid with color lightness and tonality with texture roughness. The latter found that participants preferred to associate geometrical resolution with attack time, spikiness with the spectral centroid, and visual brightness with the ratio of even to odd harmonics. Both found that a common preference among participants was much less obvious when multiple mappings were in effect.

When developing a sample browser incorporating timbre visualization, designers can either decide on implementing a fixed subset of these acoustic to visual mappings or provide options for users to modify the mappings themselves. This second option may increase the tool’s versatility but is dependant on users’ knowledge and interpretation of acoustic and visual descriptors. This is what inspired us to develop our sample browser with a simple method for users to associate textures and timbres by selecting pairs of images and samples.

2.2 Relevant work in audio browsers

Dimensionality reduction (DR) of low- and high-level audio descriptors is a common practice in audio browser research. A concise overview of commonly used DR methods in audio browsers can be found in [47] and [53]. *Islands of Music* [43] and *MusicMiner* [42] popularized using self-organizing maps (SOM) to organize music libraries into topographic maps of musical genres based on a large number of extracted low-level and high-level features. As songs are generally associated with visual metadata such as album covers and pictures of the artists, these can be used to visually differentiate and help users recognize specific songs.

CataRT [49], a tool for concatenative sound synthesis and exploration, presents sound grains (very short sound samples) in a starfield display. Originally allowing users to choose audio descriptors to define each axis and sample colors, it was later augmented with a combination of DR methods to assist sound search in large collections [48]. This tool seems to have been influential in the design of recent drum sample browsers and sequencers. *The Infinite Drum Machine* [40] demonstrated the creative possibilities of visualising t-SNE DR of drum samples in a web-based drum machine, while *XO* [60] is an example of a professional tool based on similar principles. Both tools use color and placement to differentiate sample timbre and allow users to select regions of the sample space to associate with specific beats in a rhythm sequence. Sample color is used to visualize the third dimension of the reduced space in [40], while *XO* uses sample color to distinguish predicted drum types (e.g., kick, snare, cymbal).

Stober and Nürnberger developed *MusicGalaxy* [52], a music browser proposing an innovative solution to the commonly occurring issue with dimensionality reduction that some similar elements can be projected to different regions of the reduced space. When focusing on a specific song, its nearest neighbors in the high-dimensional feature space are made obvious by increasing their size. In subsequent user studies, this method compared favorably to the more common “pan and zoom” method of navigating large collections.

The following section describes the sample browsers that are closest to ours in design that also incorporated user studies in their development.

2.3 Studies of audio sample browsers

Heise et al. [24] developed *SoundTorch*, which uses a SOM to organize environmental sound samples in 2D space. Participants preferred their method to a list-based interface. They later added a visualization of the temporal envelope as the contour of each element [25], but did not study the effects of this additional visual information on the effectiveness of the tool.

Frisson et al. [18] developed AudioMetro, which uses t-SNE DR of audio features and a proximity grid to place sound samples in a starfield display. They also use color and shape labels to differentiate samples. They map the timbre descriptor for brightness to the color lightness channel and the temporal evolution of sharpness to the contour of the visual labels. Their study mainly evaluated the effect of different methods of spatial organization of the sound samples, and they offer little analysis of the effect of the labels. They remark that simply using DR would often result in overlapping samples, which they solved by displacing samples to points on a regular grid. We encountered the same problem but implemented a different solution using simulated springs to push samples apart (Section 3.4), an approach also found in [48].

In their master’s thesis, Font [15] presented the results of queries to the Freesound database [17] using a starfield display. Their study found that participants were most successful at finding sounds when they could choose sound descriptors as axes. This was compared to placement obtained via PCA DR which they found to perform worse than random placement.

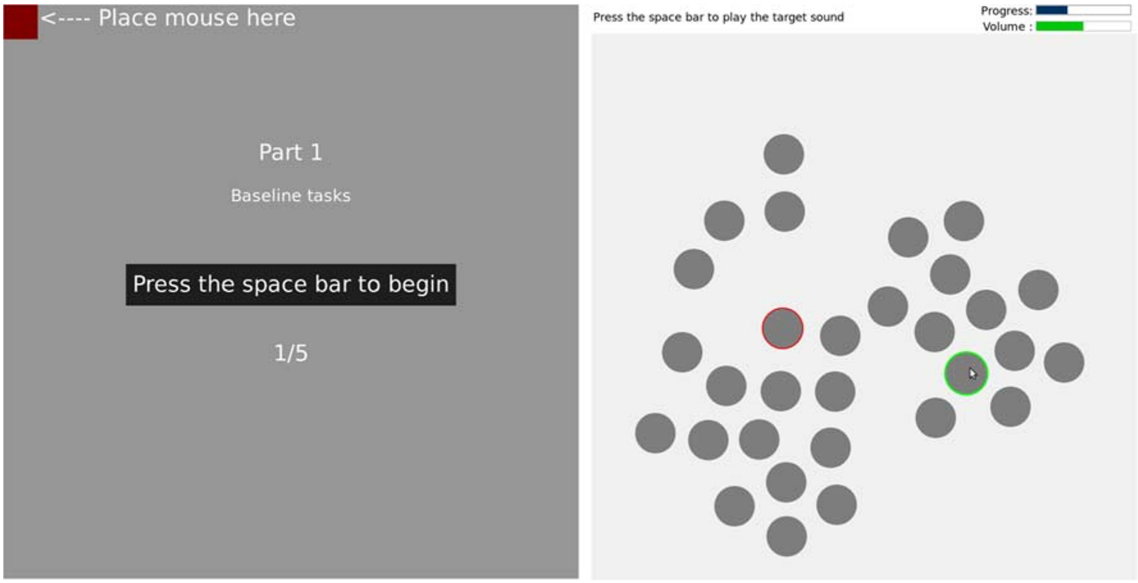
None of these studies looked in-depth into the effect of their choices of visual labels, so we made measuring this effect the main objective of our study design.

3 Study design

The goal of our study is to determine whether and to what extent different types of labels (shape, color, and texture) help in the task of searching for a specific sample in a starfield display. The secondary objective is to evaluate the effectiveness of our dimensionality reduction–based placement approach (Section 3.4). We designed a known-item search task that would allow us to test different combinations of labeling and placement methods. In order to obtain a baseline for comparison, we create baseline variants for labels and placement. The label baseline uses gray circles to represent each sample and the placement baseline assigns random coordinates to each sample.

3.1 Task design and interaction

The task interface (with baseline labels) is shown in Fig. 1. In each task, 30 sounds are picked at random from a dataset and one is designated as the target sound the participant must find. Individual sound samples are displayed as circular shapes arranged on a light gray canvas. Samples are played by mousing over the corresponding element. Pressing the space bar plays the target sound and clicking on



<---- Place mouse here

Part 1

Baseline tasks

Press the space bar to begin

1/5

Press the space bar to play the target sound

Fig. 1 The task interface used in the study. *Left*: The intermediary screen shown before each task. *Right*: An example of a baseline task. Red outlines indicate an incorrectly clicked sound while a green outline highlights the current playing sound

the correct element completes the task. Before each task, the participant must place their cursor in a corner of the canvas. This corner rotates clockwise around the canvas between tasks in order to vary the starting point.

3.2 Sets of tasks to introduce and evaluate timbre visualization methods

Participants progress through the study by completing sets of tasks that introduce and evaluate the placement method and the visual labels. They first complete a practice task with baseline labels and random placement that can be repeated until they are certain they understand how the interface works. They then complete a set of tasks with baseline labels and random placement. This represents the worst-case scenario, where no relevant information is being visualized.

The rest of the study progresses by alternating between familiarization and evaluation tasks. During familiarization tasks, participants are encouraged to take their time and explore the set of samples while searching for the target sound. During evaluation tasks, participants are instructed to find the target sound as quickly as possible. We first introduce the dimensionality reduction (DR)–based placement method with baseline labels. We then introduce a visual labeling method (color, shape, or texture) with a set of tasks that uses the DR placement. Finally, we test the effectiveness of the labels on their own in a final set of tasks with the same labeling method and random placement. Before each set of tasks, participants read some brief instructions, which can be found in the supplementary materials of the paper (Online Resource 1, Section 6).

The two placement methods (random and DR) and four label types (baseline, color, shape, and texture) combine to form 8 different testing conditions. Participants also complete three survey-style questionnaires during the study referred to as Q_0, Q_1, and Q_2. In Q_0, participants provide basic demographic information (e.g., age, listening conditions, years of musical experience). In Q_1 and Q_2, participants are asked to rate the extent to which they used different search strategies for finding the target sound. They are also asked to rate whether the positioning or labeling of the sounds helped them in their search and how difficult they found the task overall (see the supplementary materials (Online Resource 1, Section 5)) for the full list of questions.

We designed the study to introduce and evaluate each type of label individually. Table 1 summarizes how a participant would progress through the entire study. For each set of task conditions, the same task will be repeated

Table 1 Series of tasks that participants complete while progressing through the study

Order	1	2	3	4	5	6	7	8
Task type	Q_0	P	B_R	B_{DR}	L_{DR}	Q_1	L_R	Q_2

P = practice, B_x = baseline labels. L_x = color, shape, or texture labels, X_{DR}= dimensionality reduction placement, X_R= random placement

5–10 times (depending on the task) with different samples. We designed the study to take approximately 30 min to complete.

3.3 NSynth dataset and timbre feature extraction for sample placement and label generation

We use musical samples from the NSynth dataset [14], which consists of over 300,000 4-s samples produced by virtual instruments in commercial sample libraries. We are interested in differentiating timbre, so we use a subset of ∼800 samples with the same pitch and velocity. We use samples at midi note 64, which corresponds to musical note E4.

For each sample, we use a cochlear filterbank[1] [3] to extract three profiles related to the perception of timbre: a spectral envelope, a roughness envelope (which measures perceived auditory roughness over time), and a temporal amplitude envelope. These timbre features are used both to determine sample placement as well as generate visual labels.

3.4 Sample placement through dimensionality reduction

We obtain two-dimensional coordinates for the musical samples through dimensionality reduction (DR) of the extracted features. We first use PCA to reduce each profile to a shorter feature vector, then apply UMAP [41] to the concatenated vectors to produce a 2D arrangement of the samples that represents the distances between their high-dimensional timbre features. Recent related work [16, 18] has used t-SNE [35] for dimensionality reduction; however, we find that UMAP obtains comparable results in significantly less computation time.

During testing, samples would occasionally overlap in the interface. To remedy this, before displaying each task, we run a brief physical simulation by placing virtual springs between samples causing them to push away from each other if they are overlapping.

3.5 Visual label generation

Research on human visual perception has revealed that separate pathways are used to process shape, color, and texture [12]. While this indicates that shape, color, and texture could be used to distinguish between visual samples in a complementary fashion, this is not always the case. Different types of visual information can also interfere with each other [11], so we chose to test each of these pathways separately.

We are interested in differentiating timbre, which is often represented by continuous features describing some of the spectral or temporal characteristics of the sound. We generate labels that also vary continuously by using mappings from timbral features to visual parameters. Figure 2 shows how the same set of samples would appear for each type of visual label. Labels could be of varying sizes, but in our studies, their diameter was 64 pixels.

3.5.1 Shape

We use the temporal envelope to generate our shapes because it visualizes the attack time and low-frequency amplitude modulations in the signal, which are important timbral descriptors [38]. We produce a unique shape for each sound by mapping the amplitude of the temporal envelope to the contour of a circular shape. We downsample the envelope to obtain a 20-ms temporal resolution. This produces 200 distinct points for our 4-s samples. The radius of each point on a half-circle is described by (1).

$$R\left(\theta\right) = \text{envelope}[i], \quad \theta = \frac{i \cdot \pi}{200}, \quad i \in \{0, 200\} \tag{1}$$

This half-circular shape is then rotated and mirrored along the x-axis to produce a symmetrical shape. The topmost point of the shape represents the amplitude of the envelope at the beginning of the sound and the bottom-most point the amplitude at the end of the sound. This method is comparable to [25], with the main difference being that our approach produces symmetrical shapes, which are known to be easier to perceive and remember [55].

3.5.2 Color

For color, we use a simpler approach based on the coordinates of the samples in the reduced 2D space. The center of the spatial distribution of the entire set of samples is calculated and each sample is assigned a color based on its position relative to the center. The hue is determined by its angle relative to the center point and the saturation is determined by its distance to the center point. This can be imagined as laying a color wheel over the entire sample distribution and picking a color for each sample based on their location within it. Each sample's color reinforces the spatial information which is based on timbral similarity.

3.5.3 Texture

We developed a software tool to synthesize textural images for samples inspired by Li et al.'s method for "universal" style transfer [34]. The method is based on a pre-trained

[1] Available for download from https://github.com/NECOTIS/ERBlet-Cochlear-Filterbank

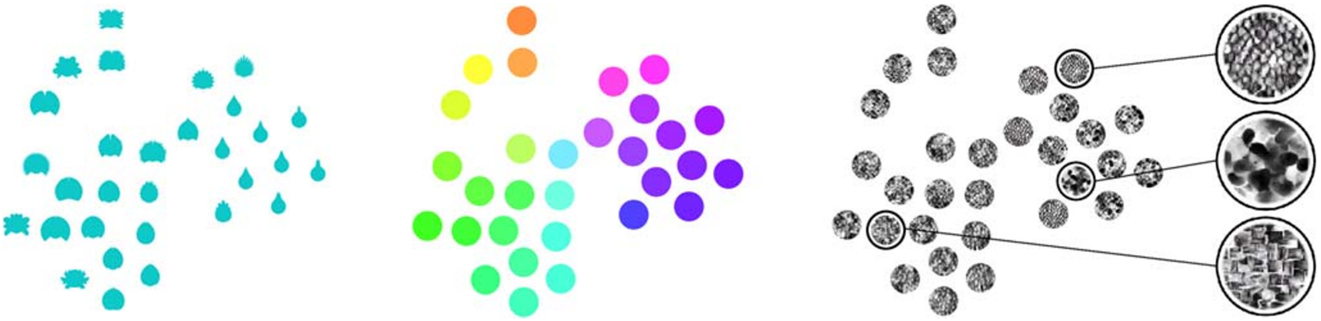


Fig. 2 The three types of visual labels for the same set of randomly selected sound samples. From left to right: shape, color, texture

encoder-decoder neural network architecture (provided by [34]). The encoder is an image classifier [50], while the decoder has been trained to reconstruct images from the activation patterns of the encoder. Noise is fed into the encoder network and the activation pattern is transformed to resemble that of a reference texture. The decoder can then produce a new image with textural properties that greatly resemble the original image. Our sample browser uses an optimized version of this architecture and provides a simple interface to extract, store, and interpolate between textural representations.

For the samples used in the study, eight medioids [27] are found in the timbre space and each one is manually assigned a texture from the normalized Brodatz texture database [1]. We use black and white texture images in order to differentiate from the color method. We choose visually distinct textures for each medioid and attempt to relate properties of the sounds to textural characteristics (e.g., a rapidly varying synth note is labeled with a chaotic rootlike texture, while a percussive mallet note is labeled with a texture of pebbles in an attempt to evoke the hardness of the material being struck). Textures for all the other sounds in the dataset are then produced by the texture synthesis method by interpolating between these textures based on their proximity to the medioids. Through this process, samples are assigned textural images whose visual properties vary in tandem with their timbral difference.

3.6 Technologies used

We use JATOS [31] to build and host our study on a webserver, which allows us to easily distribute the study to participants using a web link. JATOS studies are executed in the browser so no installation is required for the participants. The filterbank, dimensionality reduction and label generation are implemented in Python, and the resulting information is stored in a dictionary JSON file that is loaded by the web application. Pre-calculated shape and color information are stored as arrays in the dictionary. Each sample entry also points to a pre-generated JPEG texture file that is stored on the webserver. We use *jsPsych* [32] and *p5.js* [39] to build the interactive components of the study and *toxiclibs.js* [28] for the spring physics simulation.

3.7 Collected data

The measures we use to evaluate and compare labeling methods are summarized in Table 2. We collect several other data points from tasks, including the entire cursor trajectory, cursor speed, the number of misidentified samples, and the number of times the participant listened to the target sound. The anonymized data collected in our studies will be available in the supplementary materials repository [46].

In summary, we designed our study to evaluate the effect of both visual labels and placement on searching for musical samples using technologies allowing for simple and easy distribution to participants. The studies we conducted using this design were approved by the *Comité d'éthique de la recherche - Lettres et sciences humaines* of the University of Sherbrooke (ethical certificate number 2018-1795).

4 Studies

We conducted our study with 3 different groups of participants. The preliminary study did not lead to significant conclusions, prompting us to perform a study in a controlled environment with specifically qualified participants. Based on feedback from this second study, we

Table 2 Recorded measures used to evaluate task performance

Measure	Description	Units
Time	Time taken to complete the task	Seconds
Hovered samples	Number of samples the cursor encountered	Count
Distance	Total distance the mouse cursor traveled	Pixels

decided to change the manner in which color labels were generated. We conducted a third iteration of the study with two objectives in mind: testing the new color labels and comparing the quality of results obtained in controlled and uncontrolled environments.

4.1 Winter 2019: initial study

The precursor to this article [45] summarized the results from a group of 28 computer engineering students. Students completed the study as part of coursework in a class on human perception and performed the study on their own computers. They were instructed to use a computer mouse and earphones. Based on those results, we concluded that it would be worthwhile to recruit a group of participants with a minimum of 2 years musical experience to perform the study in a controlled environment. We also realized the need for the baseline task with random placement, which was originally not part of the study. We only compared completion times in this study, and decided to record more data points in follow-up studies.

4.2 Summer 2019: qualified participants in a controlled environment

Fifteen participants were recruited from the music and engineering faculties of the University of Sherbrooke. They were required to have at least 2 years of recent experience working with sound in order to qualify them for the task of differentiating sounds by timbre alone. Participants completed the study in a secluded area on tablet-style laptops with a connected mouse and keyboard. The testing stations were equipped with Sennheiser HD 280 Pro headphones, connected via a Rega EAR amplifier and a Roland UA-1G USB interface.

Participants completed three passes of the study, with each iteration testing a different labeling method. Given that there are 3 types of labels, there are 6 permutations of the order in which they could be tested. These permutations were distributed between participants as evenly as possible.

4.3 Summer 2019: new color labels and online participants

A second group of participants, recruited in the same manner as the initial study, performed the study on their personal computers. For our analysis, we used results from 14 participants who reported more than 2 years musical experience and having completed the study in good listening conditions.

We changed the way in which color labels were generated to better correspond other work in the field, particularly [22] and [2]. Our new method uses direct mapping from timbral descriptors to the hue and saturation of the circular labels. The spectral centroid (measuring timbral brightness) is mapped to a gradient from blue to red and spectral flatness (measuring tonality) is then mapped to the color's saturation. Shape and texture labels remain unchanged.

5 Analysis methods

5.1 Data transformation

Initial inspection of the collected measures (completion times, hovered samples, and total mouse cursor distance) showed that distributions were quite heavily right-tailed. After performing Box-Cox transformations [8] on each set of measures, statistical models produce normally distributed residuals.[2] Mean values and confidence intervals are calculated in transformed values and then back-transformed to their original units. Histograms of the collected measures and transformation parameters are included in the supplementary materials (Online Resource 1, Section 2).

5.2 Statistical models used to determine the effect of task conditions on performance

We use general linear mixed-effects models as they support the repeated measures that characterize our study design and account for variance between participants. The placement method and label type are modeled as fixed effects and a unique identifier assigned to each participant is modeled as a random effect. We report estimated marginal (least-squares) means and confidence intervals. Analysis of variance (ANOVA) of fitted models estimates the probability of equal means.

For significance testing in survey responses, we use the Mann-Whitney U test [36] when comparing between two groups and the Kruskal-Wallis one-way analysis of variance [29] when comparing more than two groups.

5.3 Packages and notebooks

We provide R notebooks for reproducing our results in the supplementary materials repository [46] and a list of R packages used in Appendix.

[2] For a linear regression model to be considered appropriate, the distribution of prediction errors (residuals) should resemble a normal distribution [37].

6 Results

We evaluate the effect of different labels and placement methods by comparing means of the recorded measures of task performance. We interpret *P* values under 0.05 as strong evidence that the difference between means is significant (not due to chance) [56].

6.1 Controlled study with qualified participants

We first investigate the effect of the placement method, followed by that of the labels. Finally, we present survey responses of interest.

6.1.1 Effect of placement method

Table 3 shows the mean measures grouped by placement method and by label type. In Fig. 3, we plot these means with 95% confidence intervals. We observe that for tasks with baseline labels, participants hovered 3 less samples with dimensionality reduction (DR) placement compared to random placement. Tasks with shape labels also show an ∼4 sample improvement with DR placement. For color and texture labels, while mean values of samples hovered are lower with DR placement, the difference is not statistically significant. The differences in time and distance between placement methods are insignificant for baseline tasks. For all label types aside from the baseline, the distance traveled by the mouse cursor with random placement is approximately 700 to 800 pixels longer than with DR placement. There is also a significant difference in completion times (approximately 3–4 s) in tasks with shape and texture labels.

Table 3 Means of measures grouped by placement and labeling methods. Bold *p* values indicate when the difference of mean measures between placement methods is significant

Measure	Label	Placement		*p* value
		DR	Random	
Time (s)				
	Baseline	12.3	12.7	0.54
	Shape	10.7	14.3	**0.0008**
	Color	11.6	13.7	0.051
	Texture	11.6	15.7	**0.0005**
Hovered samples				
	Baseline	14.9	17.9	**0.04**
	Shape	8.7	12.5	**0.004**
	Color	13.8	15.7	0.16
	Texture	13.5	15.3	0.30
Distance (pixels)				
	Baseline	2807	3094	0.12
	Shape	2313	3123	**0.0005**
	Color	2765	3588	**0.002**
	Texture	2681	3403	**0.009**

DR: placement by dimensionality reduction

6.1.2 Effect of labeling methods

We analyze the effect of labeling methods by comparing them to the baseline tasks. Table 4 summarizes the differences between means. Our main takeaway from these results is that with shape labels, participants need to investigate significantly fewer samples before finding the target sound. Compared to the baseline, participants visited ∼6 less samples in tasks with shape labels before finding the target sound.

Times are not significantly changed when adding visual labels, except in the case of texture labels with random placement, where participants took ∼3 s longer to find the target sound.

6.1.3 Effect of iteration

Given that participants complete the study multiple times, we are curious whether they improve at the different tasks over time. Overall, we do not see a significant effect of iteration on measures. However, in the first set baseline tasks with DR placement, there is a significant difference between completion times (p=0.0078) and mouse speed (p=0.00020) compared to subsequent iterations. We hypothesize that participants were still becoming accustomed to using the interface during this first task in the study. There are no significant differences between iterations for tasks with random placement. When inspecting the data visually across iterations, there is a noticeable downward trend for tasks with labels and DR-based placement times, but the differences do not pass significance testing.

6.1.4 Questionnaire results

The responses to many questions did not show significant differences between label types. This section highlights the most interesting responses. A full list of the questions can be found in the supplementary materials (Online Resource 1, Section 5).

Figure 4 shows Likert plots of responses to questions about the perceived consistency of the placement of samples and the labeling methods in labeled tasks with DR placement. When rating label consistency, participants are quite evenly divided on texture labels and lean towards color being more consistent. In the case of shapes, none rated their

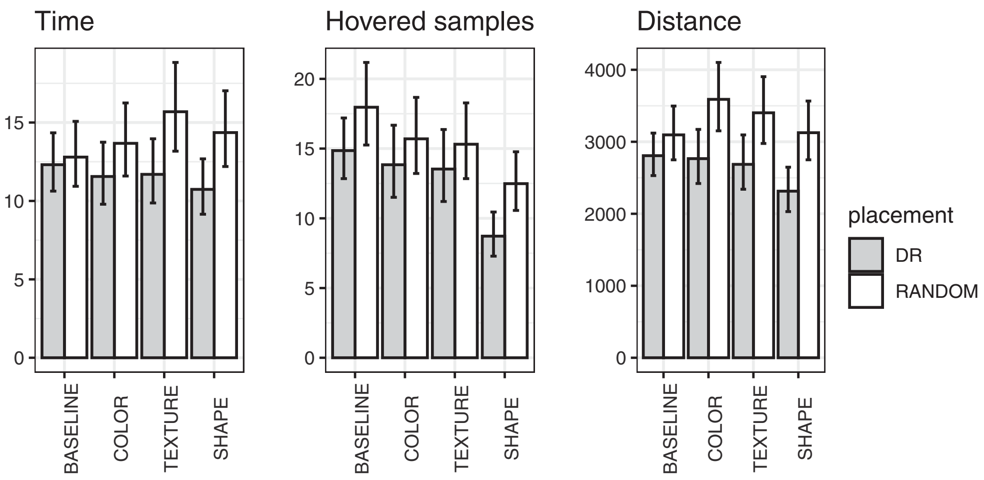


Fig. 3 Means of task measures grouped by label type and placement. 95% confidence intervals shown as error bars

consistency below 3. When asked to rate whether similar sounds were located closer together, more participants responded with lower ratings after tasks with shape labels (p=0.032).

Figure 5 shows Likert plots of ratings of the perceived helpfulness of the placement of the samples during labeled tasks with DR placement. After tasks with color and texture labels, ratings skew towards the placement being helpful, but after tasks with shape labels they rate the placement as being less helpful (p= 0.04). This indicates that participants were paying less attention to the placement when shape labels were provided. Participants are quite evenly divided between positive and negative responses when rating the helpfulness of all of the label types.

Table 4 Mean difference of measures for tasks with labels compared to baseline, grouped by placement method. Positive values indicate an improvement. Bold p values indicate that the difference from the baseline is significant

Measure	Label	Placement			
		DR		Random	
		$\overline{B} - \overline{L}$	p val.	$\overline{B} - \overline{L}$	p val.
Time (s)					
	Shape	1.59	0.18	−1.58	0.26
	Color	0.76	1.0	–0.9	1.0
	Texture	0.73	0.96	–2.91	**0.03**
Hovered samples					
	Shape	6.1	**4e−07**	5.5	**0.002**
	Color	1.1	0.89	2.3	0.52
	Texture	1.3	0.33	2.7	0.13
Distance (pixels)					
	Shape	494	**0.04**	–30	1.0
	Color	41	1.0	–493	0.20
	Texture	120	0.94	–307	0.61

$\overline{B}$: mean of measures with baseline labels,

$\overline{L}$: mean of measures with shape, color, or texture labels

DR: placement by dimensionality reduction

6.2 Online study

Participants performed this study on their own computers and were less experienced than participants in the previous study. While we used the same minimum experience criteria for both studies, 10 out of the 15 participants in the controlled environment study had at least 10 years of musical experience while 13 out of the 14 participants in this study had between 2 and 5 years of experience.

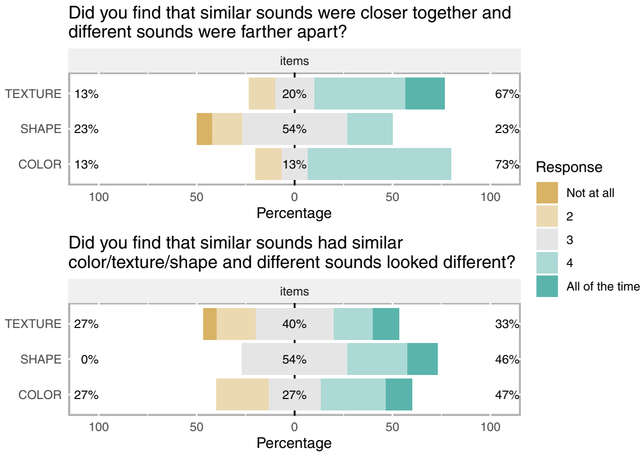


Fig. 4 Responses to the questions about the perceived consistency of the labeling and placement methods

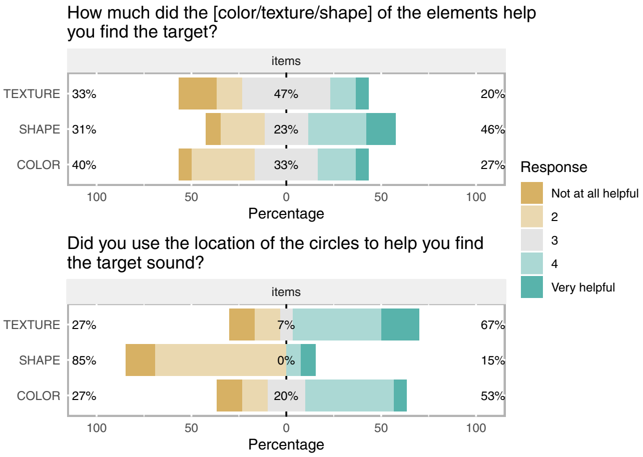


Fig. 5 Responses to the questions about the perceived helpfulness of the labels and the sample placement

6.2.1 Effect of placement methods

Figure 6 shows mean measures grouped by placement and labeling method with 95% confidence intervals. The only significant difference between placement methods is found in tasks with texture labels, where participants moved their cursor over ∼1200 more pixels searching for the target sound when positions were randomized. The exact means and p values are provided in the supplementary materials (Online Resource 1, Section 3, Table S1).

6.2.2 Effect of labeling methods

Table 5 shows the differences between labeled tasks and the baseline. There are two significant differences in completion times: on average, tasks with color labels and DR placement took 2.5 s longer than the baseline, while tasks with texture labels and random placement took 3 s longer than the baseline. Five less samples were visited on average in tasks with color labels and random placement compared to the baseline. Finally, when comparing mouse travel distance, in tasks with texture labels and random placement, participants covered ∼1100 more pixels than the baseline.

Table 5 Mean difference of measures compared to baseline, grouped by placement method. Positive values indicate an improvement. Bold p values indicate that the difference is significant

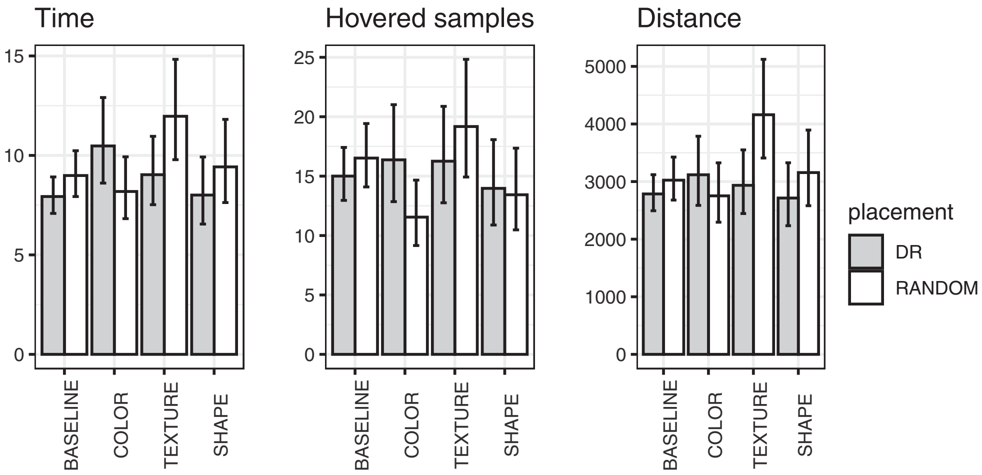
		Placement			
		DR		Random	
Measure	Label	$\overline{B} - \overline{L}$	p val.	$\overline{B} - \overline{L}$	p val.
Time					
	Shape	-0.1	1.0	-0.4	0.97
	Color	–2.5	**0.05**	0.8	0.82
	Texture	–1.1	0.56	–3.0	**0.05**
Hovered samples					
	Shape	1.0	0.96	3.1	0.50
	Color	–1.4	0.92	5.0	**0.05**
	Texture	–1.3	0.93	–2.7	0.71
Distance					
	Shape	71	0.99	–131	0.98
	Color	–333	0.72	272	0.82
	Texture	–150	0.95	–1136	**0.02**

$\overline{B}$: mean of measures with baseline labels
$\overline{L}$: mean of measures with color, texture, or shape labels
DR: placement by dimensionality reduction

7 Discussion

We designed our study to address the following two questions on the effectiveness of visual labels and sample placement in sound sample browsers.

Fig. 6 Means of measures grouped by label type and placement. 95% confidence intervals shown as error bars.



7.1 Does our timbre feature–based placement improve search, and if so, are visual labels still effective when the information provided by placement is removed?

In the controlled study with qualified participants, we see significant differences in all three measures (completion time, hovered samples, and cursor travel distance distance) when comparing our dimensionality reduction (DR) placement and random placement (Table 3). In baseline tasks, the only significant effect of sample placement was a lowering of the number of hovered samples. However, in tasks with labels, we see a significant reduction in mouse travel distance when using the DR placement. This could be explained by participants jumping between visually similar labels when positions are random, in contrast to performing a sort of nearest-neighbor or grid-like search pattern with baseline labels. We expected completion times in baseline tasks to significantly differ between placement methods, but they did not. This indicates that the tasks with random placement were easier than we expected.

As to whether labels remained effective after placement information was removed, we saw that the number of hovered samples with shape labels was ∼6 samples lower when compared to the baseline (Table 4). Additionally, the number of hovered samples is significantly lower when comparing DR and random placement within the shape label tasks (∼4 less), so we can conclude that these two methods were complementary.

In the online study, we do not see the same effect of placement methods on participant performance. This could be explained by the fact that participants completed fewer tasks with DR placement overall and may not have had enough time to learn to use the placement effectively. This somewhat contradicts our conclusion from the analysis that the effect of iteration in the controlled environment study was negligible.

7.2 Is there a difference between using color, shape, or texture as visual labels in an an audio sample browser?

In our controlled study with qualified participants, we found that shape labels significantly lowered the number of hovered samples compared to the baseline (Table 4). We interpret the reduced number of hovered sounds as an improvement in participants' ability to visually differentiate samples, and thus avoid listening to irrelevant samples. Interestingly, this gain in efficiency did not translate to a significant gain in time, which could be explained by participants spending more time visually processing the scene. We did not find any differentiation between color labels and the baseline, so we can conclude that our approach to coloring samples in this study was ineffective. In their written comments, some participants expressed that the labels produced by the color mapping were in opposition with their preconceived associations between colors and timbres. Textural labels did not differentiate from baseline tasks except in the case where participants took significantly longer to complete tasks with random placement. This indicates to us that the visual complexity of the textures was slowing them down. Many participants expressed that colored textures would have been much easier to differentiate.

The responses to the survey questions did not provide much additional insight into the differences between labeling approaches. They do however correlate with our previous observation that shape labels affected participants differently than the other two label types.

In the online study, the lower number of hovered samples in tasks with colored labels and random placement (Table 5) indicates that the new color labels could be helping participants find the target sound more effectively. This improvement is not present in the tasks with color labels and DR placement. In contrast with the previous study, shape does not seem to have much effect in reducing the number of hovered sounds when compared to the baseline.

7.3 Comparing the controlled study to the online study

Comparing both studies, the ranges of measures obtained are quite similar, but the measures from the online study have a higher variance than those from the controlled environment study (Figs. 3 and 6), making it difficult to draw many conclusions from the results.

The difference in experience noted in Section 6.2 could explain why we do not find many commonalities in the results from the two studies and suggests that our minimum criteria for experience should be raised. Given that we do not observe similar effects of visualization methods within the two groups, we are unsure whether we can recommend collecting data with this study in an uncontrolled environment. So far, our studies have been exploratory with small group sizes and our interpretations of results should be considered with this in mind. We remain optimistic that distributing this study online to a sufficient number qualified individuals would have a good chance of producing useful results.

7.4 Further work

In the studies we conducted, texture labels seemed to hinder participants more than help them. While we believe our new method for associating timbre to visual textures could prove useful for visualizing sounds in other contexts, future studies of this type of task could omit texture labels and concentrate on shapes and colors. Further work could look into whether a combination of shape and color information can outperform shape alone. A significant advantage of using color is that it is much more tolerant to large changes in scale, while shapes need to be a minimum size in order to be visually distinct.

Increasing the number of samples presented in each task would raise the overall difficulty and potentially accentuate the differences between the visualization methods being tested. We will likely also reduce the number of questions in future versions of the study, which would allow us to increase the number of tasks while maintaining the same overall duration. This study design could also be used to compare various methods of generating shapes. For example, we considered using the spectral envelope for shapes, as it contains other important timbral information such as the distribution of harmonic partials. A variant of our study worth developing would allow participants to customize each visual labeling system. This would help mitigate some issues related to visual accessibility as our current color schemes do not take color blindness into account. It would also give participants the advantage of already understanding the underlying labeling system.

8 Conclusion

We have conducted three studies using the study design presented in this article. Based on results from our first group, we decided upon minimal qualifications for participants and updated the study. Using the web based JATOS framework allowed us to quickly iterate on the study design and easily share it with our participants. Our second study brought qualified participants into a controlled environment and revealed a significant improvement when using shape labels, while texture and color labels did not provide noticeable advantages over the baseline. The final study produced few significant results, but provided some indications that our new colored labels are a step in the right direction.

Adding shape labels did not significantly improve task completion times, but did reduce the number of sounds visited before finding the correct one. It seems that the gain in efficiency of listening to less samples was offset by the extra time spent processing the additional visual information. We observed that this improvement persisted when information provided by the dimensionality reduction placement was removed, and that the two methods were complementary.

We consider our study design to have succeeded in allowing us to test a variety of placement and labeling methods and we were able to measure their individual and combined effects. We hope other researchers will use our open-source implementation of the study[3] as a starting point to pursue their own research questions related to sample browsers and sound search.

Acknowledgments We thank all of our participants for taking the time to complete our study. We also thank our reviewers for their constructive feedback. Thanks to members of the NECOTIS laboratory of the University of Sherbrooke who beta-tested the study and provided feedback. We thank CIRMMT for providing access to their research infrastucture and travel funding. We also thank Frédéric Lavoie and and the GRPA of the University of Sherbrooke for generously lending us their testing facilities. Special thanks to Felix Camirand Lemyre for his advice on statistical mode ling and analysis.

Funding information This work is partly funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the *Fonds Nature et Technologies* of Quebec (FRQNT).

Compliance with ethical standards

The studies conducted were approved by the *Comité d'éthique de la recherche - Lettres et sciences humaines* of the University of Sherbrooke (ethical certificate number 2018-1795).

Appendix: R packages

We use R [44] for our data analysis and figures. We use *forecast* [26] to estimate the optimal Box-Cox transform parameters as well as perform the forward and inverse transformations. We use general linear mixed-effect models from *lme4* [5] and *lmerTest* [30]. Estimated marginal means and confidence intervals of fitted models are calculated with *emmeans* [33]. Figures were produced with *ggplot2* [57] and *likert* [10]. *dplyr* [58] is used for data wrangling.

References

1. Abdelmounaime S, Dong-Chen H (2013) New Brodatz-based image databases for grayscale color and multiband texture analysis. Int Sch Res Not Machine Vision 2013:1–14. https://doi.org/10.1155/2013/876386
2. Adeli M, Rouat J, Molotchnikoff S (2014) Audiovisual correspondence between musical timbre and visual shapes. Front Hum Neurosci, 8. https://doi.org/10.3389/fnhum.2014.00352
3. Adeli M, Rouat J, Wood S, Molotchnikoff S, Plourde E (2016) A flexible bio-inspired hierarchical model for analyzing musical timbre. IEEE/ACM Trans Audio Speech Language Process 24(5):875–889. https://doi.org/10.1109/TASLP.2016.2530405
4. Ahlberg C, Shneiderman B Visual information seeking: tight coupling of dynamic query filters with starfield displays. In: Readings in human-computer interaction, interactive technologies. Morgan Kaufmann, pp 450–456
5. Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J of Stat Softw 67(1):1–48. https://doi.org/10.18637/jss.v067.i01

[3] Available for download from https://github.com/NECOTIS/timbre-visualisation-study

6. Berthaut F, Desainte-Catherine M, Hachet M (2010) Combining audiovisual mappings for 3D musical interaction. In: Int computer music conf. New York, USA, ICMC '10, p 9
7. Borgo R, Kehrer J, Chung DHS, Maguire E, Laramee RS, Hauser H, Ward M, Chen M (2012) Glyph-based visualization: foundations, design guidelines, techniques and applications. Eurographics 2013 - State of the Art Reports p 25 pages https://doi.org/10.2312/CONF/EG2013/STARS/039-063
8. Box GEP, Cox DR (1964) An analysis of transformations. J Royal Stat Soc Series B 26(2):211–252. http://www.jstor.org/stable/2984418, Accessed 2019-11-29
9. Brazil E, Fernstrom M (2003) Audio information browsing with the Sonic Browser. In: Proc Coord and Mult Views Conf, vol 2003, pp 26–31
10. Bryer J (2019) likert: analysis and visualization likert items. http://github.com/jbryer/likert
11. Callaghan TC (1989) Interference and dominance in texture segregation: hue, geometric form, and line orientation. Percept Psychophys 46(4):299–311
12. Cant JS, Large ME, McCall L, Goodale MA (2008) Independent processing of form, colour, and texture in object perception. Perception 37(1):57–78
13. Chen M, Floridi L (2013) An analysis of information visualisation. Synthese 190(16):3421–3438. https://doi.org/10.1007/s11229-012-0183-y
14. Engel J, Resnick C, Roberts A, Dieleman S, Norouzi M, Eck D, Simonyan K (2017) Neural audio synthesis of musical notes with wavenet autoencoders. In: Proc 34th int conf on mach learn - vol 70, JMLR.org, ICML'17, pp 1068–1077
15. Font F (2010) Design and evaluation of a visualization interface for querying large unstructured sound databases Master's thesis. Universitat Pompeu Fabra, Barcelona
16. Font F, Bandiera G (2017) Freesound explorer: make music while discovering freesound! In: Web Audio Conf. WAC 2017. London
17. Font F, Roma G, Serra X (2013) Freesound technical demo. In: Proc 21st ACM int conf on multimedia MM '13. ACM Press, Barcelona, pp 411–412. https://doi.org/10.1145/2502081.2502245
18. Frisson C, Dupont S, Yvart W, Riche N, Siebert X, Dutoit T (2014) Audiometro: directing search for sound designers through content-based cues. In: Proc of audio mostly 9 AM '14. ACM, New York, pp 1:1–1:8, https://doi.org/10.1145/2636879.2636880
19. Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks
20. Giannakis K (2006) A comparative evaluation of auditory-visual mappings for sound visualisation. Organised Sound; Cambridge 11(3):297–307
21. Grill T (2012) Constructing high-level perceptual audio descriptors for textural sounds. In: Proc. of the 9th sound and music comput. conf. (SMC 2012), Copenhagen, pp 486–493
22. Grill T, Flexer A (2012) Visualization of perceptual qualities in textural sounds. In: Int computer music conf, ICMC '12
23. Grill T, Flexer A, Cunningham S (2011) Identification of perceptual qualities in textural sounds using the repertory grid method. In: Proc 6th audio mostly conf AM '11. ACM Press, Coimbra, pp 67–74, https://doi.org/10.1145/2095667.2095677
24. Heise S, Hlatky M, Loviscach J (2008) Soundtorch: quick browsing in large audio collections. In: Proc 125th conv of the audio eng soc (2008), Paper 7544, p 8
25. Heise S, Hlatky M, Loviscach J (2009) Aurally and visually enhanced audio search with soundtorch. In: CHI '09 extended abstracts on human factors in computing systems CHI EA '09. ACM, New York, pp 3241–3246, https://doi.org/10.1145/1520340.1520465
26. Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2019) Forecast: forecasting functions for time series and linear models. http://pkg.robjhyndman.com/forecast, Accessed 2019-11-29
27. Jin X, Han J (2010) K-medoids clustering. In: Sammut C, Webb GI (eds) Enciclopedia of machine learning. Springer, Boston, pp 564–565
28. Phillips K (2011) Toxiclibs.js - open-source library for computational design. www.haptic-data.com/toxiclibsjs, Accessed 2019-11-29
29. Kruskal WH, Wallis WA (1952) Use of ranks in one-criterion variance analysis. J of the Am Stat Assoc 47(260):583–621. https://doi.org/10.1080/01621459.1952.10483441
30. Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest package: tests in linear mixed effects models. J of Stat Softw 82(13):1–26. https://doi.org/10.18637/jss.v082.i13
31. Lange K, Kühn S, Filevich E (2015) Just another tool for online studies (JATOS): an easy solution for setup and management of web servers supporting online studies. PLOS ONE 10(6):1–14. https://doi.org/10.1371/journal.pone.0130834
32. de Leeuw JR (2015) jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. Behav Res Methods 47(1):1–12
33. Lenth R (2019) emmeans: estimated marginal means, aka least-squares means. https://CRAN.R-project.org/package=emmeans, Accessed 2019-11-29
34. Li Y, Fang C, Yang J, Wang Z, Lu X, Yang MH (2017) Universal style transfer via feature transforms. Adv Neural Inf Process Syst 30:386–396
35. van der Maaten L, Hinton G (2008) Visualizing high-dimensional data using t-SNE. J Mach Learn Res 1(9):2579–2605
36. Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18(1):50–60. www.jstor.org/stable/2236101
37. Martin J, de Adana DDR, Asuero AG (2017) Fitting models to data: residual analysis, a primer. In: Hessling JP (ed) Uncertainty quantification and model calibration. IntechOpen, Rijeka, https://doi.org/10.5772/68049. chap 7
38. McAdams S, Winsberg S, Donnadieu S, De Soete G, Krimphoff J (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psych Res 58(3):177–192
39. McCarthy L (2013) p5.js | home. www.p5js.org/, Accessed 2019-11-29
40. McDonald K, Tan M (2018) The infinite drum machine. https://experiments.withgoogle.com/drum-machine, Accessed 2020-01-28
41. McInnes L, Healy J, Saul N, Grossberger L (2018) Umap: uniform manifold approximation and projection. J Open Source Softw 3(29):861. https://doi.org/10.21105/joss.00861
42. Mörchen F, Ultsch A, Nöcker M, Stamm C (2005) Databionic visualization of music collections according to perceptual distance. In: Int Soc Music Info Retrieval, ISMIR '05
43. Pampalk E, Rauber A, Merkl D (2002) Content-based organization and visualization of music archives. In: MULTIMEDIA '02
44. R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
45. Richan E, Rouat J (2019) A study comparing shape, colour and texture as visual labels in audio sample browsers. In: Proc 14th int audio mostly conf: a journey in sound, AM'19. ACM, pp 223–226, https://doi.org/10.1145/3356590.3356624
46. Richan E, Rouat J (2019) Timbre visualisation study - supplementary materials. https://doi.org/10.17605/OSF.IO/FKNHR, https://osf.io/fknhr, Accessed 2019-11-29

47. Roma G, Green O, Tremblay PA (2019) Adaptive mapping of sound collections for data-driven musical interfaces. In: New Interfaces musical expression, NIME '19
48. Schwarz D, Schnell N (2010) Sound search by content-based navigation in large databases. In: Proc 6th sound music computing conf, SMC '09
49. Schwarz D, Beller G, Verbrugghe B, Britton S (2006) Real-time corpus-based concatenative synthesis with CataRT. In: Proc 9th int conf on digital audio effects, DAFx-06
50. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv:14091556 [cs]
51. Soraghan S (2014) Animating timbre - a user study. In: Int computer music conf, ICMC '14
52. Stober S, Nürnberger A (2010) Musicgalaxy: a multi-focus zoomable interface for multi-facet exploration of music collections. In: CMMR 2010
53. Stober S, Low T, Gossen T, Nürnberger A (2013) Incremental visualization of growing music collections. In: Int Soc Music Info Retrieval, ISMIR '13
54. Walker R (1987) The effects of culture, environment, age, and musical training on choices of visual metaphors for sound. Percept Psychophys 42(5):491–502. https://doi.org/10.3758/BF03209757
55. Ward MO (2008) Multivariate data glyphs: principles and practice. In: Handbook of data vis. Springer, Berlin, pp 179–198, https://doi.org/10.1007/978-3-540-33037-0_8
56. Wasserstein RL, Lazar NA (2016) The ASA statement on p-values: context, process, and purpose. Am Stat 70(2):129–133. https://doi.org/10.1080/00031305.2016.1154108
57. Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York. https://ggplot2.tidyverse.org
58. Wickham H, François R, Henry L, Müller K (2019) dplyr: a grammar of data manipulation. https://CRAN.R-project.org/package=dplyr
59. Wolfgang K (1947) Gestalt psychology: an introduction to new concepts in modern psychology. New York, New York
60. XLN Audio (2019) XO - XLN audio. https://www.xlnaudio.com/products/xo, Accessed 2020-01-28

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.