Sketching sounds: an exploratory study on sound-shape associations

Sebastian Löbbers
Centre for Digital Music
Queen Mary University of London
s.lobbers@qmul.ac.uk

Mathieu Barthet
Centre for Digital Music
Queen Mary University of London
m.barthet@qmul.ac.uk

György Fazekas
Centre for Digital Music
Queen Mary University of London
g.fazekas@qmul.ac.uk

ABSTRACT

Sound synthesiser controls typically correspond to technical parameters of signal processing algorithms rather than intuitive sound descriptors that relate to human perception of sound. This makes it difficult to realise sound ideas in a straightforward way. Cross-modal mappings, for example between gestures and sound, have been suggested as a more intuitive control mechanism. A large body of research shows consistency in human associations between sounds and shapes. However, the use of drawings to drive sound synthesis has not been explored to its full extent. This paper presents an exploratory study that asked participants to sketch visual imagery of sounds with a monochromatic digital drawing interface, with the aim to identify different representational approaches and determine whether timbral sound characteristics can be communicated reliably through visual sketches. Results imply that the development of a synthesiser exploiting sound-shape associations is feasible, but a larger and more focused dataset is needed in followup studies.

1. INTRODUCTION

Timbre is an increasingly significant component of modern music production, often serving as a distinguishing characteristic of an artist's style or genre. Blake [1] provides an example by describing how rock bands can be divided into sub-genres by analysing timbre rather than chord progression, referencing the groups My Bloody Valentine and U2. Despite humans' capability of perceiving very subtle differences in timbre [2] it is still poorly understood by researchers [3, 4, 5]. Crafting the "right" sound is an important part of electronic music production, but synthesiser parameters typically correspond to a technical function related to signal processing, rather than a concept related to human perception, making it difficult to realise sound ideas in a straightforward way. The aim of this research is to develop a sketch-based sound synthesiser that takes simple drawings as input to provide a simple, intuitive control informed by cross-modal associations between sounds and shapes. The development will be informed by the results of a study, presented in this paper, that explores how participants represent timbre with a digital drawing interface. This section introduces relevant research into sound-shape associations and presents related work about sound synthesis control. Sections 2, 3 and 4 describe the design, analysis and results of the study. The discussion in Section 5 is followed by a conclusion in Section 6.

1.1 Background on Sound-Shape Associations

Strong evidence suggests that a majority of people think of sound in a visual way to some extent that references colour, brightness, shapes and contour [6]. In the 1920s, Köhler discovered that humans associate the made-up word *takete* with sharp, jagged shapes, and *maluma* with soft, round shapes [7]. Similar findings were made with the words *boubou* and *kiki* [8] and generally among all phonemes [9]. This effect has been observed across cultures [10, 11, 12], age groups including toddlers [13] and, to some extent, with the visually impaired [14]. Adeli et al. [15] and Grill et al. [16] found similar associations between shapes and musical instruments or abstract sonic textures respectively. Focusing on pitch, loudness and tempo, Küssner et al. [17] asked participants to draw their representations of sound rather than selecting existing shapes. They found that representations are not only influenced by musical structure, but also by a participant's music proficiency. Engeln and Groh [18] loosely classified drawings of sounds that were re-synthesised from spectrograms into real-world associations like scenes, actions or emotions, abstract shapes and structures or references to audio visualisations like waveforms and amplitude envelopes.

1.2 Related Work

A variety of strategies has been proposed to improve the intuitiveness of sound synthesisers. Low-level synthesis parameters can be mapped to a smaller number of high-level descriptors that either correspond to concepts of human perception as implemented in the *FeatSynth* framework [19] or describe the actions and objects that produce a sound, as seen with the impact sound synthesiser developed at the PRISM laboratory [20, 21]. Further, a different input modality like gestures [22] or voice [23] allows for a more intuitive control, potentially affording the simultaneous manipulation of multiple parameters. This can be particularly helpful for the exploration and drafting of sound ideas. Timbre visualisations, as mentioned in section 1.1, are more frequently used in a sound retrieval context [24], but have also been adapted for synthesis, for example by *Sound Mosaic* [25] which allows users to manipulate shapes to drive sound synthesis. Knees and Andersen [26] explored how drawings could be used for sound retrieval with a non-functioning prototype, an

approach more commonly found in image search applications [27, 28]. A major challenge of developing a sketch-based sound synthesiser is to find meaningful mappings between visual features and synthesis parameters.

2. METHODS AND MATERIAL

This section describes the design of an exploratory study that investigates how participants represent sound stimuli intuitively through free-form sketching with a digital drawing interface. The following hypotheses were put forward:

- There will be some level of agreement between participants on how to sketch a sound.
- Correlations between quantitative sketch and audio features will align with the sound-shape associations described in Section 1.1.
- A participant's music proficiency and the sound type will have an influence on the representational approach, but overall abstract sketches will be produced more frequently than depictions of real-world associations.

2.1 Participants

Twenty-eight participants were recruited through mailing lists and in person at the School of Electronic Engineering and Computer Science at Queen Mary University of London. This group was divided equally by gender (14 female, 14 male), 25 were adults below the age of 33 (three between 34 and 49), 22 had a Western background (16 Europe, 4 North America, 2 South America) and 5 an Eastern background (4 China, 1 India) with one participant preferring not to disclose this information. As described in Section 2.4, participants were divided by music proficiency resulting in 14 musicians and 14 non-musicians.

2.2 Stimuli

A total of ten timbrally dissimilar sounds were selected following the research of Adeli et al. [15] and Grill et al. [16] ranging from musical instruments (*Piano*, *Strings*, *Electric Guitar*) and environmental sounds (*Impact*) to synthesised pads (*Telephonic*, *Subbass*) and abstract textures (*Noise*, *String Grains*, *Crackles*, *Processed Guitar*). [1] All sound stimuli are monophonic, normalised for equal loudness, pitched to the MIDI note C3 and last eight seconds including trailing silence to mark a clear endpoint during looped playback. The perceived base frequency may vary due to prominent harmonics.

2.3 Apparatus

The drawing interface was implemented in *p5.js* and runs in a web-browser. White strokes can be drawn on a 750x750 pixel canvas with a black background that separates it from the rest of the page. Stroke colour or width cannot be changed and sketches cannot be modified or erased. This design was chosen to encourage participants to follow their intuition and focus on shape rather than visual texture and colour. Clicking and dragging the mouse cursor starts a sketch and the timestamped cursor position is recorded consecutively while sketching. The study was conducted in person using the trackpad on a 15" MacBook Pro and a pair of Beyerdynamic DT 770 Pro headphones in calm, indoor locations.

2.4 Procedure

Participants first completed a questionnaire that included an excerpt of the Gold MSI framework [29] and was used to determine their music proficiency. [2] Participants were asked to familiarise themselves with the drawing interface without audio before they were presented with the sound stimuli. The study intended to encourage a spontaneous response, therefore no information about the range of sounds was provided and participants were instructed to sketch what they believed to represent each sound stimulus the best. Looped playback started automatically with the option to pause/resume. Each sound was played twice in a randomised order resulting in a total of twenty sketches per participant. After completion, a short semi-structured interview was conducted and audio recorded asking participants how they approached the task and whether they found it difficult. No time limit was given and the study typically took twenty to thirty minutes to complete. The study setup can be accessed online. [3]

3. ANALYSIS

This section describes the qualitative and quantitative methods used to analyse the collected data.

3.1 Interview Analysis

The interview analysis aimed to identify and summarise different approaches to the task and quantify reported difficulty. Interviews were transcribed and thematic analysis [30] was used to find reoccurring themes. Task difficulty was coded into hard/neutral/easy depending on the response to the question, "How difficult did you find the task?".

3.2 Sketch Categorisation

Section 2 introduced the hypothesis that participants will predominately produce abstract sketches. This can be tested by dividing sketches into high-level categories that refer to their representational approach. In order to minimise bias, sketches were categorised in an open card sorting study by 6 participants (4 female, 3 musicians) who did not take part in the main study. They were asked to create a reasonable number of categories (three to ten was recommended) and write a short description for each of them. The study was completed remotely within three hours on participants' devices. [4] Results were reduced to two dimensions with principal component analysis (PCA) and clustered with the K-Means algorithm. The silhouette coefficient [31], a measure of cluster goodness, was calculated to find the most suitable number of clusters between three and ten. Clusters were annotated and named with the

[1] All sounds can be accessed online together with the sketches drawn by participants during the experiment `https://bit.ly/3ta6crU`.

[2] A participant was categorised as a musician if they scored above average and reported involvement in musical activity.

[3] Study setup: `https://bit.ly/3j3FkVO`

[4] Card sorting instructions: `https://youtu.be/LXTlnaAciWw`

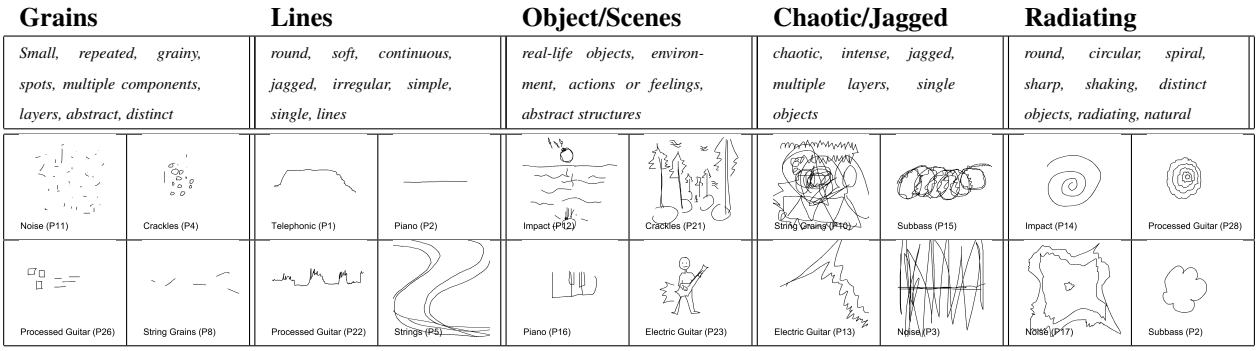

Table 1. Sketch categories with examples. Category names and keywords were obtained through thematic analysis as described in Section 3.2. *Objects/Scenes* mainly refers to real-world associations while other categories highlight different abstract approaches, but category clusters might overlap with a number of sketches showing characteristics of more than one category. Colours were inverted for better visibility.

help of keywords that were extracted from participants' descriptions using thematic analysis.

3.3 Sketch Feature Extraction

While sketch categories give an overview of the representational approaches, a more detailed, quantitative set of features is needed to compare sketches in detail and find correlations with sound characteristics using statistical analyses described in Section 3.5. A number of features can be calculated from the sketches' data shape and through simple arithmetic operations as demonstrated in Equations 1, 2 and 3, where N is the number of strokes in a sketch and $\overline{L}$, $\overline{T}$ and $\overline{S}$ are their average length, completion time and drawing speed. The total number of points in the k^{th} stroke is described by n_k. Each point has a position x_{k_i} and timestamp t_{k_i}. The euclidean distance between two points is described by $d(p, q)$.

$$\overline{L} = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=2}^{n_k} d\left(x_{k_i}, x_{k_{i-1}}\right) \tag{1}$$

$$\overline{T} = \frac{1}{N} \sum_{k=1}^{N} t_{k_{n_k}} - t_{k_1} \tag{2}$$

$$\overline{S} = \frac{1}{N} \sum_{k=1}^{N} \sum_{i=2}^{n_k} d\left(x_{k_i}, x_{k_{i-1}}\right) \frac{1}{t_{k_{n_k}} - t_{k_1}} \tag{3}$$

Sound-shape associations are usually reported with respect to a shape's contour focusing on their "jaggedness" or "roundness" [15, 16]. These attributes were quantified by extracting corner points divided into obtuse, right and acute angles [32] and curve points divided into wide and narrow shape [33]. A qualitative review suggested that sketches differ by the number of stroke intersections that can be interpreted as the "noisiness" of a sketch. The number of intersections was determined using an adaptation of Bresenham's rasterisation algorithm [34]. Prior to extracting features, the sketch data was cleaned by removing consecutive points with the same position and merging two strokes if a starting point was within a five pixel distance to an end point. The number of intersections, corner and curve points is reported relative to the total stroke length of a sketch. [5]

[5] A detailed summary of audio and sketch feature extraction can be found at `http://doi.org/10.5281/zenodo.4764351`.

3.4 Audio Feature Extraction

In order to investigate sound-shape associations through statistical analysis, the sound stimuli also have to be described with quantitative features. This was accomplished by computing the mean values of *Centroid Frequency*, *Spectral Flatness*, *Zero Crossing* and *Root Mean Square Power (RMS)* [35] for each sound using the *Librosa* library with a FFT window size of 2048 and hop length of 512. In addition, the timbre models proposed by Pearce et al. [36] provided quantified measures of *Hardness*, *Depth*, *Brightness*, *Roughness*, *Warmth*, *Sharpness* and *Boominess*. The additional feature *RMS Slope*, describing how continuous or intersected a sound is, was quantified by the slope between prominent extrema in the RMS envelope.

3.5 Statistical Analyses

Differences in sketch category counts between participant groups and between sounds were computed using Pearson's Chi-squared test and Cochran's Q test respectively. Spearman's rank coefficient was used to find significant correlations between audio features and mean sketch features between all participants. To determine whether inter-rater reliability of sound descriptions can be measured with the sketch features introduced in Section 3.3, the ICC(2,k) model [6] of the intraclass correlation coefficient (ICC) [37] was deployed. Sketch features were first log-transformed to meet the normal distribution assumption of the ICC.

4. RESULTS

This section presents the results of the interview analysis, sketch categorisation, inter-rater reliability testing and correlation between audio and sketch features.

4.1 Interview

Task difficulty was reported as easy/hard/neutral by 14/8/6 participants. Participants who felt that the task was easy thought that "there was no right or wrong" (P4), "it was just about being creative" (P15) and they did not have to "achieve something" (P10). On the contrary, others found it difficult to "think of sound in a very visual way" (P8)

[6] For the ICC, sound stimuli were defined as subjects and sketch features as measurements. Repeated sounds were considered separate subjects.

(P16). While some participants approached the task as an intuitive, creative activity, others were concerned with establishing a consistent visual language. Difficulties arose while deciding which sound characteristics to follow because "there are too many things to consider" like "brightness or aggressiveness or how it [timbre] develops over time" (P6). A consistent approach was difficult to maintain because of a "great variety in the sounds" (P2). Some participants reported that "complicated ones [sounds] sounded like pictures, and then the simple ones [...] like piano notes were a lot harder to draw" (P8) possibly because they "hear [them] all the time" (P1), while other participants thought that "it's pretty straightforward because I know a piano note more than others" (P5).

4.2 Sketch Categories

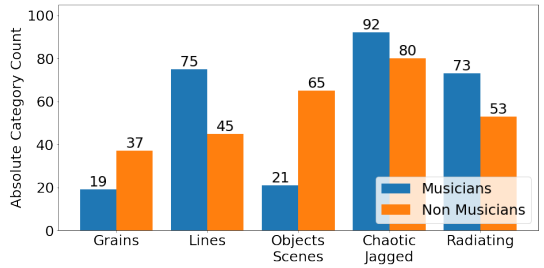

Figure 1. Absolute category counts by music proficiency

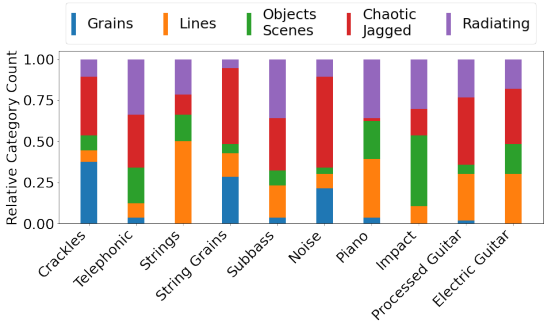

Figure 2. Relative category counts by sound stimulus

Analysis of the card sorting study described in Section 3.2 returned an optimal number of five categories that were named: *Chaotic/Jagged* (172 sketches), *Radiating* (126), *Lines* (120), *Objects/Scenes* (86) and *Grains* (56). Descriptive keywords and sketch examples for each category can be found in Table 1. A maximal silhouette coefficient of 0.49 suggests that categories are distinguishable, but not clearly separated which is also reflected by occasionally overlapping keywords. Chi-squared test suggests that non-musicians produce *Objects/Scenes* sketches more often ($\chi^2(1,N=28)=22.51$ $p<.0001$) while musicians produce *Lines* sketches at a higher rate ($\chi^2(1,N=28)=7.5$ $p<.01$) possibly because this category contains sketches that appear to reference audio visualisations like envelopes or waveforms. Category counts for *Objects/Scenes* sketches significantly differ between sounds ($\chi^2(9)=67.07$ $p<.0001$) with post-hoc analysis revealing that *Piano* and *Impact* show significantly higher counts than *Noise*, *String Grains* and *Processed Guitar* ($p<.01$ for each pair).

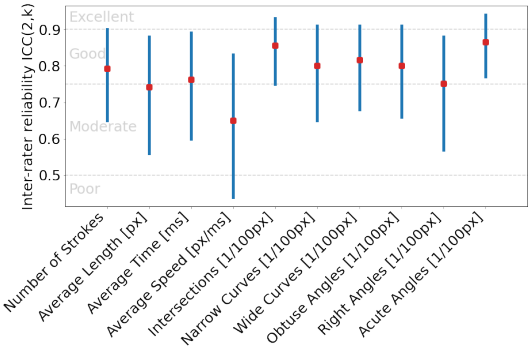

Figure 3. Mean values and 95% CI of ICC(2,k) inter-rater reliabilities for each sketch feature with evaluation guidelines proposed by Koo and Li [37] ($df1=19$, $df2=513$, $p<.01$ for all features)

4.3 Inter-rater Reliability

As seen in Figure 3, reliability measures were good to excellent for *Intersections* and *Acute Angels*, poor to good for *Average Speed* and moderate to good for all remaining features within the 95% confidence interval (CI) suggesting that some level of agreement exists between participants on how to represent sounds visually and that it can be measured with the sketch features introduced in Section 3.3.

4.4 Feature Correlations

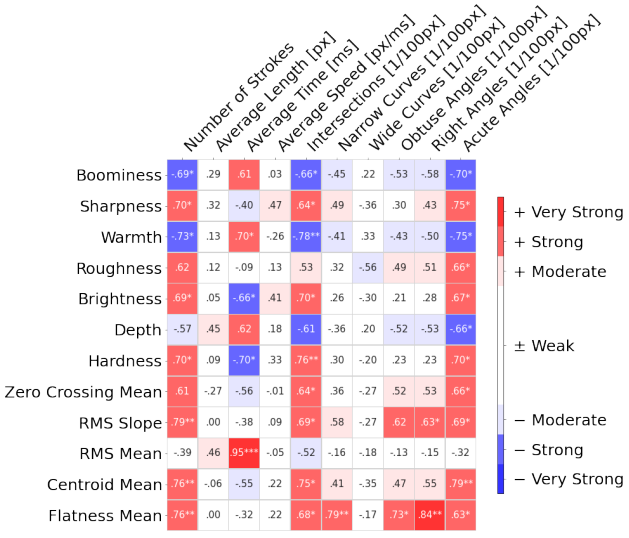

Figure 4. Spearman's rank correlation coefficients between sketch and audio features with annotated p-values: $p<.05$ (*), .01 (**), .001 (***)

Several significant correlations were found between sketch and audio features. *Acute Angles* (11), *Intersections* (9) and *Number of Strokes* (8) show the highest statistically significant ($p<.05$) number of strong ($r>.6$) and very strong ($r>.8$) correlations with audio features. The strongest correlation overall was found between *RMS Mean* and *Average Time* ($r=.95$, $p<.001$). Opposing audio features like *Warmth* and *Sharpness* showed similar absolute correlation values but opposite directions for *Number of Strokes*, *Intersections* and *Acute Angles*.

5. DISCUSSION

All participants completed the study successfully, but the exploratory study design described in Section 2.4 benefited participants who approached the task intuitively and made it more difficult for those who aimed to follow a consistent, systematic approach. The card sorting results provide meaningful high-level categories that should, however, not be interpreted as mutually exclusive with many sketches showing elements of more than one category. Overall, participants predominately chose abstract elements like shapes and contours over scenes or icons, but this does not necessarily capture a participant's intention. For example, a sketch showing a single line might visualise the simplicity of a sound, but could also refer to its amplitude envelope. Music proficiency and the type of sound had an influence on the representational approach with depictions of real-life associations being more prevalent among non-musicians and for instrumental or environmental sounds as reported in Section 4.2. Some participants reported having experience with digital drawing applications and, while this was not quantified in the study, it does appear to have had an effect on their approach. A different interface, like a graphics tablet, might also have an impact on the results. The ICC analysis suggests that some level of agreement exists between participants on how to sketch sounds visually and that these sketches can be described reliably with the features introduced in Section 3.3. However, the ICC(2,k) model used in the analysis considers the averaged measure of all raters and cannot provide information about the reliability of a single rater. A sketch-based synthesiser needs to work with the input of individual users one at a time and therefore rely on measurements with a high single rater reliability. In future work, the suitability of the sketch features has to be evaluated in that context. A large number of strong correlations between sketch and audio features was found that generally align with results from studies where existing visualisations were matched to sounds [15, 16]. Sharp, rough and hard sounds result in sketches with more acute angles compared to warm or deep sounds. Contrary to expectation, curve points did not show any significant correlations as shown in Figure 4. This could either mean that warm sounds were represented with lines rather than curves, which is supported by significant negative correlations between *Warmth/Depth* and *Intersections* or indicate that a shape's roundness was not represented well by the curve points. *Number of Strokes* and *Intersections*, features not commonly discussed in sound-shape research, produced strong correlations and should be considered in future work. A qualitative review led to the hypothesis that *Objects/Scenes* sketches will not produce the same correlations as abstract sketches because they correspond to high-level associations rather than specific sound characteristics. However, subsets were too small for conclusive quantitative evaluation. Generally, a larger dataset would produce more robust results for all statistical tests.

6. CONCLUSION

In this exploratory study, a variety of possible user responses were found that could be expected when implementing a sketch-based sound synthesiser that uses a digital drawing interface. Results indicate that there is a general consensus about how to communicate timbre through visual sketches that can be quantified in a statistically meaningful way by extracting visual and audio features. These findings support the assumption that the development of a sketch-based synthesiser is feasible. However, an exploratory study design can only provide general, indicative results. Future work will have to focus on a specific set of parameters on which a cross-modal mapping paradigm can be built. Stricter instructions, a larger sample size, possibly focusing on a specific type of participant, and a smaller, targeted set of sound stimuli, for example using only synthesised pads, could be beneficial to produce more detailed results.

Acknowledgments

EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1).

7. REFERENCES

[1] D. K. Blake, "Timbre as differentiation in indie music," *Music Theory Online*, vol. 18, no. 2, 2012.

[2] H. Peng and J. D. Reiss, "Why Can You Hear a Difference between Pouring Hot and Cold Water? An Investigation of Temperature Dependence in Psychoacoustics," in *Audio Engineering Society Convention 145*. Audio Engineering Society, 2018.

[3] C. Saitis, S. Weinzierl, K. von Kriegstein, S. Ystad, and C. Cuskley, "Timbre semantics through the lens of crossmodal correspondences: A new way of asking old questions," *Acoustical Science and Technology*, vol. 41, no. 1, pp. 365–368, 2020.

[4] K. Siedenburg, I. Fujinaga, and S. McAdams, "A comparison of approaches to timbre descriptors in music information retrieval and music psychology," *Journal of New Music Research*, vol. 45, no. 1, pp. 27–41, 2016.

[5] S. McAdams and B. L. Giordano, "The perception of musical timbre," *The Oxford handbook of music psychology*, pp. 72–80, 2009.

[6] G. Martino and L. E. Marks, "Synesthesia: Strong and weak," *Current Directions in Psychological Science*, vol. 10, no. 2, pp. 61–65, 2001.

[7] W. Köhler, "Gestalt Psychology.[Psychologische Probleme 1933]," *New York Horace Liveright*, 1929.

[8] V. S. Ramachandran and E. M. Hubbard, "Synaesthesia–a window into perception, thought and language," *Journal of consciousness studies*, vol. 8, no. 12, pp. 3–34, 2001.

[9] A. K. S. Nielsen and D. Rendall, "Parsing the Role of Consonants versus Vowels in the Classic Takete-Maluma Phenomenon." *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, vol. 67, no. 2, pp. 153–163, 2013.

[10] R. Davis, "The Fitness of Names to Drawings. a Cross-Cultural Study in Tanganyika," *British Journal of Psychology*, vol. 52, no. 3, pp. 259–268, 1961.

[11] I. K. Taylor and M. M. Taylor, "Phonetic Symbolism in Four Unrelated Languages," *Canadian Journal of Psychology/Revue canadienne de psychologie*, vol. 16, no. 4, pp. 344–356, 1962.

[12] A. J. Bremner, S. Caparos, J. Davidoff, J. de Fockert, K. J. Linnell, and C. Spence, ""Bouba" and "Kiki" in Namibia? A remote culture make similar shape–sound matches, but different shape–taste matches to Westerners," *Cognition*, vol. 126, no. 2, pp. 165–172, 2013.

[13] D. Maurer, T. Pathman, and C. J. Mondloch, "The Shape of Boubas: Sound–Shape Correspondences in Toddlers and Adults," *Developmental Science*, vol. 9, no. 3, pp. 316–322, 2006.

[14] R. Bottini, M. Barilari, and O. Collignon, "Sound symbolism in sighted and blind. The role of vision and orthography in sound-shape correspondences," *Cognition*, vol. 185, pp. 62–70, 2019.

[15] M. Adeli, J. Rouat, and S. Molotchnikoff, "Audiovisual correspondence between musical timbre and visual shapes," *Frontiers in human neuroscience*, vol. 8, p. 352, 2014.

[16] T. Grill and A. Flexer, "Visualization of Perceptual Qualities in Textural Sounds," . *As*, p. 8, 2012.

[17] M. B. Küssner, D. Tidhar, H. M. Prior, and D. Leech-Wilkinson, "Musicians Are More Consistent: Gestural Cross-Modal Mappings of Pitch, Loudness and Tempo in Real-Time," *Frontiers in Psychology*, vol. 5, 2014.

[18] L. Engeln and R. Groh, "CoHEARence of audible shapes—a qualitative user study for coherent visual audio design with resynthesized shapes," *Personal and Ubiquitous Computing*, pp. 1–11, 2020.

[19] M. Hoffman and P. R. Cook, "The Featsynth framework for feature-based synthesis: Design and applications," in *ICMC*, 2007.

[20] A. Bourachot, K. Kanzari, M. Aramaki, S. Ystad, and R. Kronland-Martinet, "Perception of the object attributes for sound synthesis purposes," in *Computer Music Multidisciplinary Research 2019*, 2019.

[21] M. Aramaki, M. Besson, R. Kronland-Martinet, and S. Ystad, "Controlling the Perceived Material in an Impact Sound Synthesizer," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 301–314, Feb. 2011.

[22] P. Esling, N. Masuda, and A. Chemla–Romeu-Santos, "FlowSynth: Simplifying Complex Audio Generation Through Explorable Latent Spaces with Normalizing Flows," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020, pp. 5273–5275.

[23] I. Ekman and M. Rinott, "Using vocal sketching for designing sonic interactions," in *Proceedings of the 8th ACM conference on designing interactive systems*, 2010, pp. 123–131.

[24] E. Richan and J. Rouat, "A Proposal and Evaluation of New Timbre Visualisation Methods for Audio Sample Browsers," *Personal and Ubiquitous Computing*, 2020.

[25] K. Giannakis, "A Comparative Evaluation of Auditory-Visual Mappings for Sound Visualisation," *Organised Sound*, vol. 11, no. 3, pp. 297–307, 2006.

[26] P. Knees and K. Andersen, "Searching for audio by sketching mental images of sound: A brave new idea for audio retrieval in creative music production," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 2016, pp. 95–102.

[27] P. Sousa and M. J. Fonseca, "Sketch-Based Retrieval of Drawings Using Spatial Proximity," *Journal of Visual Languages & Computing*, vol. 21, no. 2, pp. 69–80, 2010.

[28] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-Based Image Retrieval by Salient Contour Reinforcement," *IEEE Transactions on Multimedia*, vol. 18, no. 8, pp. 1604–1615, 2016.

[29] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, "The musicality of non-musicians: an index for assessing musical sophistication in the general population," *PloS one*, vol. 9, no. 2, p. e89642, 2014.

[30] V. Braun and V. Clarke, "Using Thematic Analysis in Psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[31] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[32] A. Wolin, B. Eoff, and T. Hammond, "ShortStraw: A Simple and Effective Corner Finder for Polylines." in *SBM*, 2008, pp. 33–40.

[33] Y. Xiong and J. J. LaViola Jr, "Revisiting shortstraw: improving corner finding in sketch-based interfaces," in *Proceedings of the 6th Eurographics Symposium on Sketch-Based Interfaces and Modeling*, 2009, pp. 101–108.

[34] J. E. Bresenham, "Algorithm for computer control of a digital plotter," *IBM Systems journal*, vol. 4, no. 1, pp. 25–30, 1965.

[35] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," *CUIDADO Ist Project Report*, vol. 54, no. 0, pp. 1–25, 2004.

[36] A. Pearce, T. Brookes, and R. Mason, "Modelling Timbral Hardness," *Applied Sciences*, vol. 9, no. 3, p. 466, 2019.

[37] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.