SketchSynth: cross-modal control of sound synthesis

Sebastian Löbbers, Louise Thorpe, and György Fazekas

Centre for Digital Music, Queen Mary University of London, United Kingdom
{s.lobbers, l.thorpe, george.fazekas}@qmul.ac.uk

Abstract. This paper introduces a prototype of *SketchSynth*, a system that enables users to graphically control synthesis using sketches of cross-modal associations between sound and shape. The development is motivated by finding alternatives to technical synthesiser controls to enable a more intuitive realisation of sound ideas. There is strong evidence that humans share cross-modal associations between sound and shapes, and recent studies found similar patterns when humans represent sound graphically. Compared to similar cross-modal mapping architectures, this prototype uses a deep classifier that predicts the character of a sound rather than a specific sound. The prediction is then mapped onto a semantically annotated FM synthesiser dataset. This approach allows for a perceptual evaluation of the mapping model and gives the possibility to be combined with various sound datasets. Two models based on architectures commonly used for sketch recognition were compared, convolutional neural networks (CNNs) and recurrent neural networks (RNNs). In an evaluation study, 62 participants created sketches from prompts and rated the predicted audio output. Both models were able to infer sound characteristics on which they were trained with over 84% accuracy. Participant ratings were significantly higher than the baseline for some prompts, but revealed a potential weak point in the mapping between classifier output and FM synthesiser. The prototype provides the basis for further development that, in the next step, aims to make *SketchSynth* available online to be explored outside of a study environment.

Keywords: Sound synthesis control · Sound sketching · Cross-modal mapping · Musical timbre perception · Deep learning · Sketch recognition · Human-computer interaction

1 Introduction

Digital technology is now ubiquitous in music production. Eliminating the need for expensive analogue equipment and studio space enables a larger number of people to produce music. As a result, the sound of contemporary music is fundamentally shaped by digital synthesisers, audio effects and sample libraries. However, these tools are typically organised in reference to technical concepts,

which can make it difficult to realise sound ideas in a straightforward way. Recent developments seek to close this gap by centering their designs around human perception, often with the help of machine learning. The aim of this research is to develop *SketchSynth*, a system that allows for the exploration of a synthesiser space by sketching one's visual association with sound. The design is informed by multiple studies that asked participants to sketch their sound associations. The results show that similarities in sketched representations exist between participants, but individual human factors introduce significant noise to the data, a common challenge in cross-modal research. When asking participants to sketch a sound it cannot be determined with certainty which characteristic primarily influenced their representation. For example, a sound that could be described semantically as *noisy* and *thin* might be represented with focus on only one of these descriptors. This raises the question whether sound-sketches show greater similarity among participants if they are produced while imagining a sound from a semantic description, rather than listening to sounds directly. The work presented in this paper follows two objectives: first to implement and evaluate a proof-of-concept prototype of *SketchSynth* and second to collect a dataset of sketches that were produced to semantic prompts describing a sound. The results provide the basis for future work that will develop *SketchSynth* to be tested in a music practice context.

The paper is structured as follows: Section 2 introduces related work in music production tools and outlines relevant research about sound-shape associations and sketch recognition; Section 3 describes the design of *SketchSynth* and evaluation methods followed by the results, discussion and conclusion in Sections 4, 5 and 6.

2 Related Work

Many recent developments that aim to simplify music production with the help of artificial intelligence can be summarised under the umbrella term Intelligent Music Production (IMP) [29]. IMP research is increasingly implemented into commercial software; for example *XO* by XLN audio[1] is based on perception-informed re-organisation of sample libraries for easier sound exploration and retrieval [5,13], or Izotope's mixing and mastering plugins[2] which build on research into automatic and assisted mixing [8]. Other works explore different modes of interaction, for example, retrieving synthesiser or audio effect parameters from sounds or vocal mimicry [32,12,28], and synthesis control through gestures [37] or visual sound metaphors [14]. By implementing a functioning prototype for the first time, this research extends the proposal for a sketch-based sound retrieval tool by Knees and Andersen [20]. Through interviews with music producers, they found that mental concepts of sound are often rooted in the visual domain.

[1] `https://www.xlnaudio.com/products/xo`
[2] `https://www.izotope.com/en/shop/mix-master-bundle-advanced.html`

2.1 Sound-shape associations

People frequently reference visual concepts like colour, brightness, shapes and contour when they think of sound [26]. Associations between sound and shapes were first described by the connections that humans make between the made-up word pairs, *maluma/takete* or *bouba/kiki*, and jagged and round shapes [21,30]. This effect was observed across cultures [7,33,4], age groups including toddlers [27], the visually impaired [2], and between shapes and musical instruments [1] or abstract sounds [15]. Recent studies asked participants to sketch their personal associations with sound rather than using existing visual stimuli [24,10,22]. The results can be categorised into figurative and abstract representations, with the latter showing correlations between visual and sound features that align with prior sound-shape research. An evaluation study showed that participants can successfully match these abstract representations and their corresponding sounds [25]. Engeln et al. trained an end-to-end autoencoder that can be used for sketch-based sound query [11]. A similar approach is deployed in this work, however here a real-time sketch input is mapped to sounds using a model that is trained to predict sound characteristics.

2.2 Sketch recognition

Sketch recognition is typically used in the context of image retrieval but can also find application for cross-modal mapping tasks like *SketchSynth*. The established approach using convolutional neural networks (CNNs) significantly outperforms conventional machine learning methods at the benchmark task of classifying handwritten digits with the MNIST dataset [23,9]. Increasingly popular are recurrent neural networks (RNNs) that can take advantage of the sequential vector format in which digital sketches are typically saved. Seminal work by Ha and Eck [16] introduced the *Quick, Draw!* dataset and the Sketch-RNN architecture for sketch classification and generation. *Quick, Draw!*[3] is a large open-source dataset with over 50 million sketches that enables researchers to experiment and pre-train models for specific tasks. While a RNN classifier can already outperform a CNN on complex sketch classes by learning temporal relationships, CNNs might be more suited for learning abstract visual structures [36]. While *SketchRNN* produces impressive results for sketch classification and generation, algorithmic approaches might be more suitable for describing the shape of a sketch. Wolin et al.'s *ShortStraw* algorithm [34] provides a simple, effective tool to extract corner points. Xiong et al. [35] extended the algorithm to also recognise curve points. Szegin et al. [31] further show that information can not only be extracted from a sketch's shape but also from the drawing speed.

3 Methods and Material

This section describes the design and evaluation of the *SketchSynth* prototype. A deep learning approach is used to predict sound characteristics from a sketch

[3] `https://quickdraw.withgoogle.com/`

input, which is then mapped onto an annotated FM synthesiser dataset. Two architectures, a CNN and a RNN, were compared and a study was conducted to obtain participant ratings of the predicted sounds. The study was designed as a between-subject multivariate test where two different model architectures were tested against a random baseline. The following hypotheses were postulated:

- Sketches produced from a semantic prompt describing a sound can achieve higher accuracy than sketches produced from a sound stimulus.
- A binary classifier can be trained to distinguish between sketches of two sound classes above the random baseline of 50% accuracy.
- A mapping model using this binary classifier will receive significantly higher ratings from human participants compared to the random baseline.
- The CNN and RNN architectures will perform similarly for classification accuracy and participant ratings.

3.1 Sketch dataset

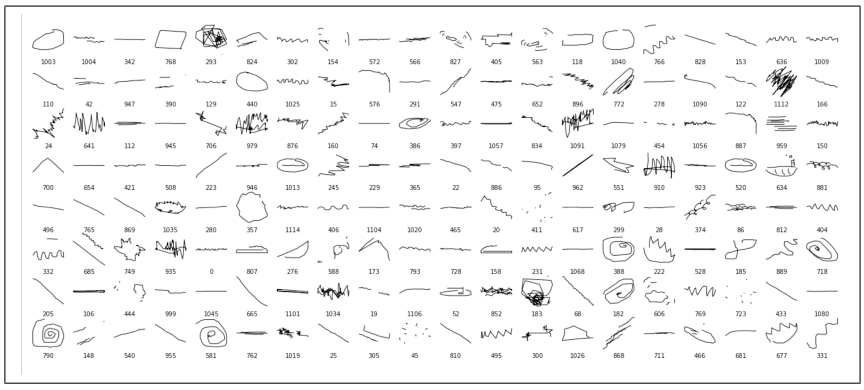

Fig. 1. A representative subset of sketches from the *Sketching Sounds* dataset. The dataset can be accessed online at `https://doi.org/10.5281/zenodo.7590916`.

Both classifiers were trained on the *Sketching Sounds* dataset that contains 1760 sound-sketches collected in a study with 88 participants that followed the design of an earlier study by the author[24]. Participants sketched their association with synthesiser sounds described in Section 3.2 with a digital interface similar to the one shown in Figure 5. In addition, each participant created two sketches from a prompt (*Draw a calm/noisy sound*) in a pre-task test without audio. While larger sketch datasets exist as discussed in Section 2.2, *Sketching Sounds* is the only dataset of sketched synthesiser sound representations to date.

3.2 Sound dataset

The *Sketching Sounds* dataset was created on a subset of 20 sounds sourced from a dataset by Hayes et al. [19] that includes 364 synthesiser sounds. Each sound was saved as a set of parameters that belong to a browser-based FM synthesiser.[4]This dataset was chosen because it includes annotations from participants that enable perceptual analysis of the sound-sketches. Thirty music producers created the sounds from prompts (*bright*, *rough*, *thick*), which were semantically rated by a subset of 24 English speaking participants along a scale of 30 sound descriptors [17,18]. Factor analysis of these annotations found five semantic factors (*sharpness*, *mass*, *clarity*, *percussiveness*, *rawness*) that explained 74% of data variance.

3.3 Sketch-to-sound mapping using deep learning

As shown in Figure 2, the sketch-to-sound mapping consists of two parts: (1) a binary classifier predicting the sound category from a sketch input and (2) the selection of a suitable sound from the FM synthesis dataset described in Section 3.2. This simple architecture was chosen to allow for a transparent, perceptual interpretation. In addition, this modular setup makes it possible to easily connect a sketch-input to a different set of sounds that was annotated by humans or by an automated music-information retrieval approach. For the binary classi-

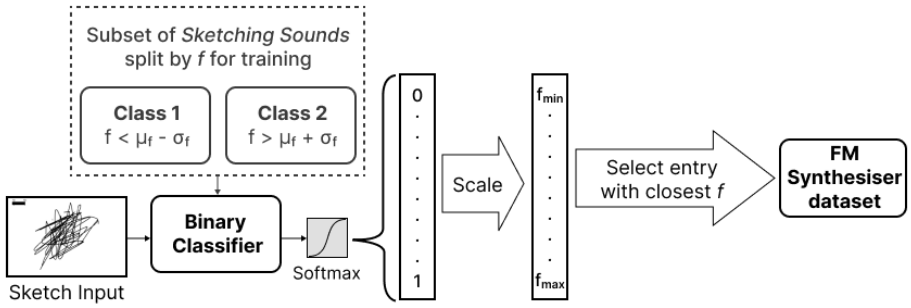

Fig. 2. Mapping logic between sketch input and sound output. A binary classifier predicts the sound-category from two opposing classes (e.g. *noisy* and *calm*) which is then used to pick a suitable sound from an annotated FM synthesiser dataset. Multiple binary classifiers can be trained to simultaneously predict multiple sound categories. A demonstration of the setup is available at `https://youtu.be/ca1LYn8Yy-g`.

fication, six subsets were extracted from the *Sketching Sounds* dataset described in Section 3.1. For each of the five semantic factors described in Section 3.2 a

[4] The synthesiser implementation can be accessed at `https://github.com/ben-hayes/fm-synth-study`

subset was created by first calculating the mean rating of that semantic factor and then sorting sketches corresponding to sounds with a rating one standard deviation below or above the mean into two classes. The test sketches produced to the semantic prompts *calm* and *noisy* described in Section 3.1 formed an additional subset.

Two deep classification architectures commonly used for sketch recognition tasks, as discussed in Section 2.2, were compared: a CNN and a RNN. Both models were implemented in Keras with cross-entropy as the loss function and accuracy as the evaluation metric. To address the relatively small sample size, sketches were augmented through rotation, scaling, dropouts and Perlin noise applied to sketch points [6,16]. In addition, the RNN architecture was pre-trained on the geometric categories *squiggle*, *zigzag*, *square*, *triangle*, *circle* and *line* from the *Quick, Draw!* dataset that represent similar abstract structures as the *Sketching Sounds* dataset. Figure 3 shows the RNN architecture that uses the encoder part of the Sketch-RNN variational autoencoder (VAE)[5] proposed by Ha and Eck [16] for feature extraction and three dense layers for classification. After pre-training, the feature-extractor layers were frozen to accelerate the training time for the sound-sketches. Figure 4 shows the CNN architecture that uses a structure commonly used for image classification like MNIST handwritten digit classification [9]. This architecture was chosen because *Sketching Sounds* contains simple, monochromatic representations similar to MNIST. In addition, a simple model might provide faster predictions when used in a real-time, client-side setup for future work, as discussed in Section 5. The classifiers' softmax output was scaled to the range of the semantic factor ratings in the FM dataset to retrieve the sound with the closest annotated value. For the *calm* and *noisy* subset, the output was scaled to the ratings of the *noisy* descriptor, where a negative rating represents a *calm* sound. For the random baseline, synthesiser sounds were picked randomly from the FM dataset after a sketch input was received.

3.4 Participants

Sixty-two participants were recruited internally at the author's institution and externally through the ISMIR[6] mailing list. Twenty-two identified as female, 37 as male, 2 as other and one participant preferred not to disclose this information. Ages ranged from to 22 to 58 ($\mu = 30.16$, $\sigma = 6.66$). The majority of participants (46) work in engineering, computer science or psychology, some with a focus on music. Only three were outside of these fields. Thirteen described themselves to be engaged in academia, either as students, PhD candidates, postdocs or academics, without specifying their field. Survey responses indicate a high level of music experience with median responses showing engagement in musical activity multiple times a week, actively listening to music 60-90 minutes per day and 6-9

[5] Keras implementation of Sketch-RNN used in this paper can be accessed at `https://github.com/KKeishiro/Sketch-RNN`

[6] `https://ismir.net/`

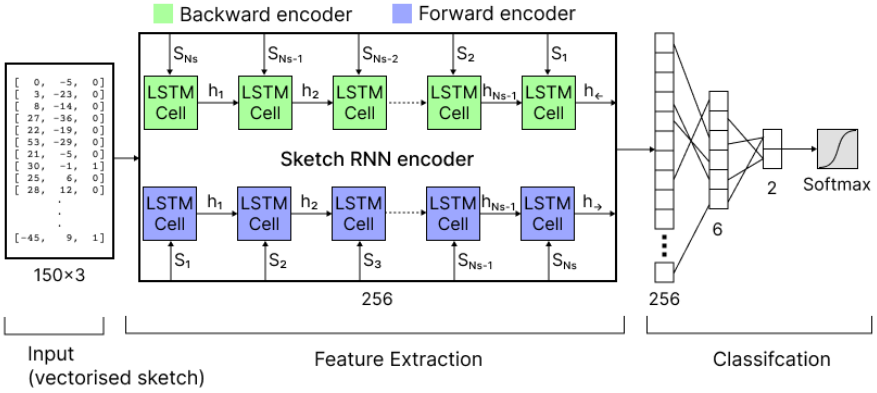

Fig. 3. Classification model using the encoder architecture from Sketch-RNN [16]. The feature extraction part of the network was pre-trained on a subset of the *Quick, Draw!* dataset as explained in Section 3.3.

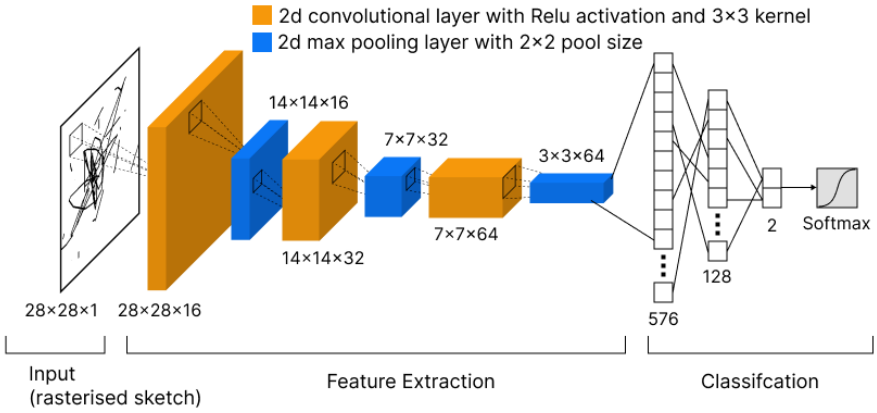

Fig. 4. CNN Classification model.

years of formal music education. Relating to experience with the visual arts, these responses were considerably lower, with median responses showing engagement in art creation for 0 hours in a typical week, consuming art multiple times per year and 0 years of formal education in a visual art or design discipline.

3.5 Apparatus

According to the methods outlined in Section 4.1, the best performing models for the RNN and CNN architectures were selected for the participant study. The models were deployed using a *Flask* backend. The study used the digital interface shown in Figure 5. Each round started with a prompt displayed in the middle of the canvas that faded out after starting a sketch. Finished sketches were sent to the backend to be saved in a database and to predict a suitable sound. The sound parameters were then returned to the participant who rated the synthesised sound on a scale from 0 (no match) to 100 (perfect match).

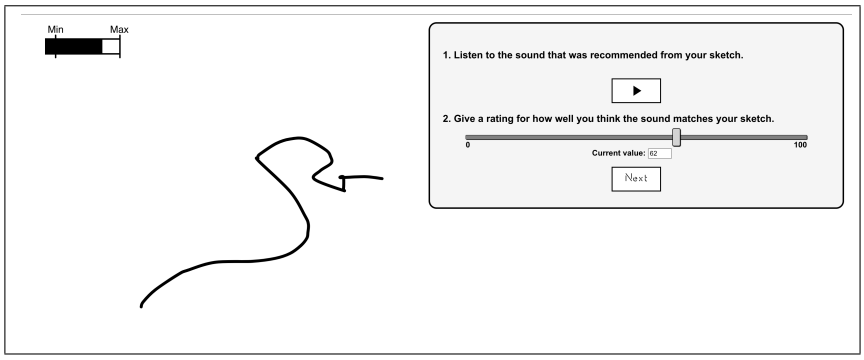

Fig. 5. Screenshot of the online study interface that was. A similar interface was used to collect the *Sketching Sounds* dataset described in Section 3.1. To encourage simple, abstract representations, the sketch length was limited to a range visualised by a meter in the top left corner of the canvas.

3.6 Procedure

Participants were first presented with a set of information ensuring that they use a laptop or desktop and were able to listen to sound either through headphones or loudspeakers. This was followed by a short introduction of the study with guidelines on representing sounds in an abstract rather than figurative way, a short explanation of the sketching interface and a guide to adjusting playback volume to a comfortable level. The main task consisted of six rounds in randomised order in which participants sketched according to a prompt and rated the resulting sound as explained in Section 3.5. The prompts were structured as

Draw a [descriptor] sound using the descriptors: *calm*, *noisy*, *clean*, *rough*, *thin*, *thick*. The descriptors were derived from the FM synthesis study by Hayes et al. [19] introduced in Section 3.2. *Bright/dark* were replaced with *noisy/calm* in this study, because, as described in Section 4.1, the models trained on the *calm/noisy* subset were chosen for participant evaluation. The study concluded with a survey collecting information about participants' demographic data, experience with music and art, hardware that was used to complete the study and feedback about their overall experience.

4 Analysis and results

This section describes the evaluation of the classifiers and the analyses and results of participant ratings and evaluation. All sketches collected in the evaluation study can be accessed at `https://doi.org/10.5281/zenodo.7591067` together with sound predictions and participant ratings.

4.1 Model evaluation

K-fold cross-validation with 10 folds was used to evaluate the deep classifiers. For each run, one fold was used as a test set and the remaining nine formed the train and validation set with a 90-10 split. As shown in Table 1, both architectures performed best on the *calm/noisy* subset which was consequently chosen for the models in the participant evaluation study. Evaluated on the sketches produced in that study, they returned accuracies of 84.21% for RNN and 92.31% for CNN. These results are higher than the average performance of both architectures on K-fold validation sets. Detailed predictions can be seen in Figure 6.

	CNN		**RNN**	
Subset	Mean Acc. [%]	Std. Dev. [%]	Mean Acc. [%]	Std. Dev. [%]
calm/noisy	**71.82**	**9.80**	**73.41**	**10.86**
SF1 (Sharpness)	52.66	3.81	51.74	4.30
SF2 (Mass)	60.23	3.50	64.06	7.48
SF3 (Clarity)	53.81	3.53	54.48	2.70
SF4 (Percussiveness)	60.72	3.05	64.18	5.88
SF5 (Rawness)	55.35	6.54	62.38	2.68

Table 1. Mean accuracies and standard deviations for binary classifiers trained on sketch subsets described in Section 3.3 and evaluated with 10 fold cross-validation explained in Section 4.1. Semantic Factors (SF) described in Section 3.2 are presented with name suggestions by Hayes et al. [18] in parenthesis.

4.2 Sound ratings

A main objective of this study was to find out whether satisfactory sound predictions could be made with a simple, generalised mapping model. To test for

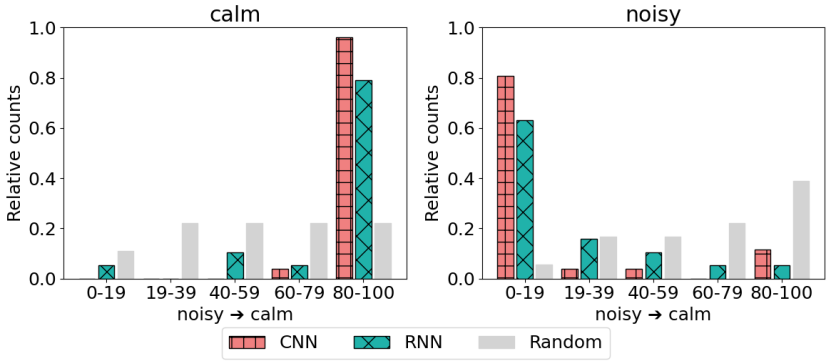

Fig. 6. Prediction histograms for the *calm* and *noisy* prompts. The x-axis of each subplot corresponds to the softmax output of the *calm/noisy* classification model with 0 referring to a completely *noisy* and 100 to a completely *calm* prediction.

significant differences, the Mann-Whitney U-test was used to compare between ratings for the random baseline model and the CNN and RNN models that were trained on the *calm/noisy* subset. Figure 7 shows the ratings participants gave for the *calm* and *noisy* prompt with annotated significance levels in comparison to the random baseline. CNN and RNN both received significantly higher ratings for the *calm* prompt ($p<.01$ for CNN and $p<.05$ for RNN); however, no significant difference could be found for the *noisy* prompt. Interestingly, CNN predictions were rated significantly higher for *clean* despite not being trained on this semantic class. This could be explained with correlations between the sound descriptors *calm* and *clean* that were also found by Hayes et al. [18].

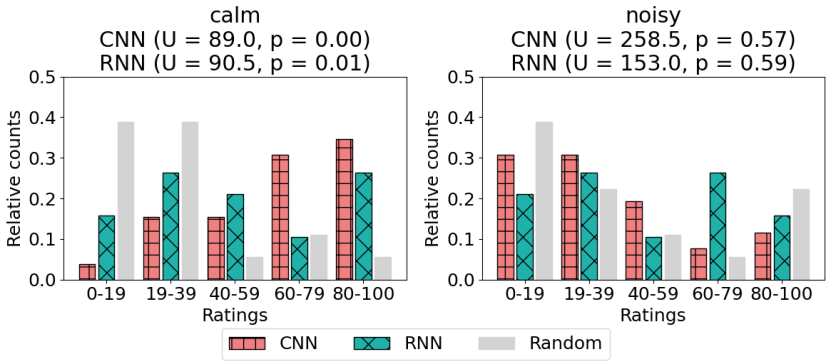

Fig. 7. Rating histograms for classification models and random baseline. Statistics and p-values are reported for Mann-Whitney U test between distributions of the respective model and the random baseline.

4.3 Survey responses

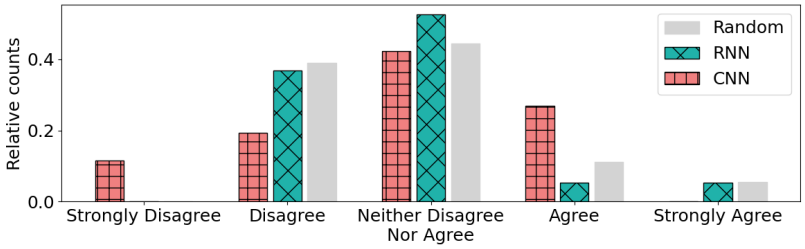

Fig. 8. Answer to system evaluation question: *I thought that this system produced suitable sounds from my sketches.*

Participants' overall experience with the system was collected on a five point Likert scale and presented separately for each mapping scheme as shown in Figure 8. Distributions look similar for CNN, RNN and the random baseline with most participants giving neutral ratings for the system. Responses skewed slightly negative for random (82.35% said *neither agree/disagree* or *disagree*) and RNN (89.47% said *neither agree/disagree* or *disagree*) and slightly positive for CNN (69.23% said *neither agree/disagree* or *agree*). Significant differences to the random baseline were not expected for the overall experience as the classification model was only trained to recognise *calm* and *noisy* sketches. General feedback was submitted as free-form text and summarised with thematic analysis [3] to identify common remarks. Multiple participants found that sound predictions were either too similar or that the same sound was played multiple times: "[...] there was not much change between them. I encountered 3 different sounds basically [...]" (P16); "I think one sound incorrectly played twice" (P22); "Not much variance in the proposed sound - had few times the same ones." (P42). This could be confirmed quantitatively from the sound predictions showing that most participants who were presented with a prediction model heard a repeated sound at least once during the study (13 of 19 for RNN and 20 of 26 for CNN), compared to only 1 of 17 for the random baseline. Some participants criticised the predictions stating that, "the sound didn't match my sketches" (P1), or "the sound I imagined based on the description was often entirely different to the sound I heard" (P33). Others found the predictions to be accurate: "pure sine tone from the circle I drew was a nice mapping, the mapping from jagged lines to rough sounds was also pleasing" (P12); "The calm sounds work well, as they are pure tones" (P42). These comments were all made by participants who were presented with either the CNN or RNN. Overall, multiple positive comments about the study were left that suggest a wider interest in the system: "This was really interesting. I enjoyed using the system a lot." (P61); "Anyways I loved the experience!" (P31).

4.4 Evaluation of semantic sound-sketches

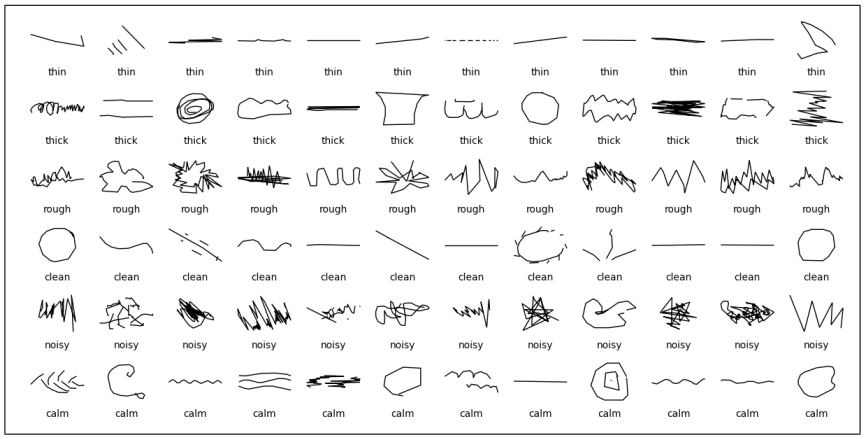

Fig. 9. A representative selection of sketches that were collected in this study. Each row shows sketches that were produced to a different semantic prompt.

The model evaluation in Section 4.1 suggests that using a sound-sketch dataset based on semantic prompts can be classified with higher accuracy than a dataset produced from sound stimuli. This hypothesis was further investigated with the semantic sound-sketch dataset presented in Figure 9 that was collected in this study. A multi-class deep classifier was trained using the RNN architecture visualised in Figure 3 that achieved better classification results than the CNN, as shown in 1. The encoder part used for feature extraction remained the same, but the classification part was changed to two fully connected layers each with 128 units and a 6-dimensional Softmax output corresponding to the number of semantic descriptors used in the study. The model was pre-trained on the *Quick, Draw!* subset described in Section 3.3. The encoder was frozen for training with the semantic sound-sketch dataset with a 84-16 training-test split. The results of the evaluation with 72 test sketches shown in Figure 10 show significantly higher accuracy than the random baseline of 16.6% (2 in 12 correct predictions) for all prompts except for *thick*. The results further suggest similarities between *noisy* and *rough* sketches with both classes being most often misclassified as the respective other. Surprisingly, *calm* and *clean* sketches were not confused with one another as suggested in Section 4.2 but were most often misclassified as *thin*. Predictions for *thick* are spread across multiple classes, however qualitatively assessing *thick* and *thin* sketches in Figure 9 suggests that a binary classifier might be able to distinguish between these opposing classes.

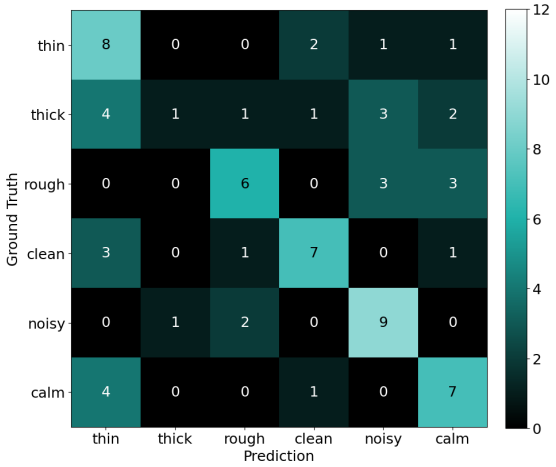

Fig. 10. Confusion matrix for the multi-class classifier that was trained on sketches created to semantic prompts. The test set consisted of 12 sketches for each prompt.

5 Discussion

The results of the classification model presented in Section 4.1 indicate that sound-sketches can be distinguished more easily with a deep learning model when they are produced to semantic prompts like *Draw a calm/noisy sound*. It is not always clear which sound characteristics participants represented in their sketches when listening to a sound, which can lead to a larger variance in representation compared to using prompts. However in Section 4.4, the evaluation of the semantic sound-sketch dataset collected in this study suggests participants do not represent all perceptual sound dimensions with distinctly different sketch approaches. Similar approaches can be found between classes; in Figure 9, for example, circular shapes are seen for *thick*, *clean* and *calm*. However, these shapes do not appear in the opposing classes *thin*, *rough* and *noisy*. While this hypothesis needs to be investigated systematically, it does hint that a model using multiple binary classifiers between opposing semantic classes (*noisy*/*calm*, *clean*/*rough*, *thick*/*thin*) might be able to achieve higher accuracies and would enable the simultaneous prediction of multiple perceptual classes. Section 4.1 suggests that the Sketch-RNN architecture does not provide a significant advantage over the simpler CNN architecture for this specific use. This might be due to the abstract structure of the sketches which are less complex than many of the *Quick, Draw!* categories and might, therefore, not require a complex architecture. Contrary to expectation, both models achieved higher classification accuracies with the sketches collected in this study compared to the *Sketching Sounds* dataset. This might be resulting from the different study design or participant pool that drew primarily from a population with high music expertise. Comparing Figures 6 and 7 show that, while predictions for *noisy* and

calm prompts were highly accurate, participants' ratings of the suggested sounds were less positive, implying a bottleneck in the mapping between deep learning model output and synthesiser parameters. This is reflected in some of the qualitative feedback presented in Section 4.3 with participants noticing similar or same sounds being produced multiple times. This can be explained with the behaviour of classification models that push their output to 0 or 1 which, following the mapping model shown in Figure 2, leads to sounds with the maximal or minimal annotated value for *noisy* being selected more often. These results show that a single, generalised mapping architecture can predict suitable sounds to some level; however, the performance can be improved through a number of approaches: (1) parameters of the current classification models could be fine-tuned for this specific task or the architecture could be improved, for example by combining RNNs and CNNs; (2) the current prototype only returns a sound after a sketch is submitted. A design that returns sounds while sketching would provide immediate feedback to a user allowing them to adjust their sketches accordingly and continuously explore the synthesiser sound space; (3) the softmax output of the classifier could be interpreted as a relative change (e.g. increasing the value for *noisy* to select a sound), which would prevent the over-representation of a small number of sounds; (4) the modular mapping architecture could be replaced with an end-to-end model that directly predicts synthesis parameters; however, this would make a perceptual interpretation difficult; (5) predictions could adjust to personal preferences, for example through reinforcement learning. The *SketchSynth* prototype presented in this work fulfilled its function as a proof-of-concept, but future work will need to move away from a perceptual research environment to evaluate the concept in a music practice context. A next step could invite music producers to explore the system and reflect on how they might integrate it into their practice.

6 Conclusion

A first prototype of *SketchSynth* was implemented successfully and evaluated in a participant study. The results show that suitable sounds can already be predicted with a simple, generalised mapping architecture. The prompt-based sound-sketch dataset collected from the study provides a basis for extending the prediction model to additional perceptual categories. Participant feedback and quantitative analyses of the architectures will inform the future development of this system. The next step seeks to release a readily-available online version of *SketchSynth* that will allow for interaction outside of a study environment.

7 Acknowledgements

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

References

1. Adeli, M., Rouat, J., Molotchnikoff, S.: Audiovisual correspondence between musical timbre and visual shapes. Frontiers in human neuroscience **8**, 352 (2014)
2. Bottini, R., Barilari, M., Collignon, O.: Sound symbolism in sighted and blind. the role of vision and orthography in sound-shape correspondences. Cognition **185**, 62–70 (2019)
3. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative Research in Psychology **3**(2), 77–101 (2006). https://doi.org/10.1191/1478088706qp063oa
4. Bremner, A.J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K.J., Spence, C.: "bouba" and "kiki" in namibia? a remote culture make similar shape–sound matches, but different shape–taste matches to westerners. Cognition **126**(2), 165–172 (2013)
5. Bruford, F., Barthet, M., McDonald, S., Sandler, M.B.: Groove explorer: An intelligent visual interface for drum loop library navigation. In: IUI Workshops (2019)
6. Das, A., Yang, Y., Hospedales, T., Xiang, T., Song, Y.Z.: Sketchode: Learning neural sketch representation in continuous time. In: International Conference on Learning Representations (2021)
7. Davis, R.: The Fitness of Names to Drawings. a Cross-Cultural Study in Tanganyika. British Journal of Psychology **52**(3), 259–268 (1961). https://doi.org/10.1111/j.2044-8295.1961.tb00788.x
8. De Man, B., Reiss, J., Stables, R.: Ten years of automatic mixing (2017)
9. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine **29**(6), 141–142 (2012)
10. Engeln, L., Groh, R.: Cohearence of audible shapes—a qualitative user study for coherent visual audio design with resynthesized shapes. Personal and Ubiquitous Computing pp. 1–11 (2020)
11. Engeln, L., Le, N.L., McGinity, M., Groh, R.: Similarity analysis of visual sketch-based search for sounds. In: Audio Mostly 2021, pp. 101–108 (2021)
12. Esling, P., Masuda, N., Chemla–Romeu-Santos, A.: FlowSynth: Simplifying Complex Audio Generation Through Explorable Latent Spaces with Normalizing Flows. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. pp. 5273–5275 (2020). https://doi.org/10.24963/ijcai.2020/767
13. Garber, L., y Ciencia, M.A., Ciccola, T., Amusategui, J.C.: Audiostellar, an open source corpus-based musical instrument for latent sound structure discovery and sonic experimentation. In: Proceedings of ICMC (2020)
14. Giannakis, K.: Sound Mosaics: A Graphical User Interface for Sound Synthesis Based on Audio-Visual Associations. Ph.D. thesis, Middlesex University (2001)
15. Grill, T., Flexer, A.: Visualization of perceptual qualities in textural sounds. In: International Computer Music Conference (ICMC). p. 8 (2012)
16. Ha, D., Eck, D.: A Neural Representation of Sketch Drawings. arXiv:1704.03477 [cs, stat] (May 2017)
17. Hayes, B., Saitis, C., Fazekas, G.: Perceptual and semantic scaling of fm synthesis timbres: Common dimensions and the role of expertise. ICMPC-ESCOM (2021)
18. Hayes, B., Saitis, C., Fazekas, G.: Disembodied timbres: A study on semantically prompted fm synthesis. Journal of the Audio Engineering Society **70**(5), 373–391 (2022)
19. Hayes, B., Saitis, C., et al.: There's more to timbre than musical instruments: semantic dimensions of fm sounds (2020)

20. Knees, P., Andersen, K.: 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. pp. 95–102 (2016)
21. Köhler, W.: Gestalt psychology.[psychologische probleme 1933]. New York Horace Liveright (1929)
22. Küssner, M.B., Tidhar, D., Prior, H.M., Leech-Wilkinson, D.: Musicians are more consistent: Gestural cross-modal mappings of pitch, loudness and tempo in real-time. Frontiers in Psychology **5** (2014). https://doi.org/10.3389/fpsyg.2014.00789
23. LeCun, Y.: The mnist database of handwritten digits. http://yann. lecun. com/exdb/mnist/ (1998)
24. Löbbers, S., Barthet, M., Fazekas, G.: Sketching sounds: an exploratory study on sound-shape associations. In: International Computer Music Conference (ICMC). p. 6 (2021)
25. Löbbers, S., Fazekas, G.: Seeing sounds, hearing shapes: a gamified study to evaluate sound-sketches. In: International Computer Music Conference (ICMC). p. 6 (2022)
26. Martino, G., Marks, L.E.: Synesthesia: Strong and weak. Current Directions in Psychological Science **10**(2), 61–65 (2001)
27. Maurer, D., Pathman, T., Mondloch, C.J.: The shape of boubas: Sound–shape correspondences in toddlers and adults. Developmental Science **9**(3), 316–322 (2006). https://doi.org/10.1111/j.1467-7687.2006.00495.x
28. Mehrabi, A., Dixon, S., Sandler, M.B.: Vocal imitation of synthesised sounds varying in pitch, loudness and spectral centroid. The Journal of the Acoustical Society of America **141**(2), 783–796 (Feb 2017). https://doi.org/10.1121/1.4974825
29. Moffat, D., Sandler, M.B.: Approaches in intelligent music production. In: Arts. vol. 8, p. 125. Multidisciplinary Digital Publishing Institute (2019)
30. Ramachandran, V.S., Hubbard, E.M.: Synaesthesia–a window into perception, thought and language. Journal of consciousness studies **8**(12), 3–34 (2001)
31. Sezgin, T.M.: Feature Point Detection and Curve Approximation for Early Processing of Free-Hand Sketches p. 77
32. Singh, S., Bromham, G., Sheng, D., Fazekas, G.: Intelligent control method for the dynamic range compressor: A user study. Journal of the Audio Engineering Society **69**(7/8), 576–585 (2021)
33. Taylor, I.K., Taylor, M.M.: Phonetic symbolism in four unrelated languages. Canadian Journal of Psychology/Revue canadienne de psychologie **16**(4), 344–356 (1962). https://doi.org/10.1037/h0083261
34. Wolin, A., Eoff, B., Hammond, T.: Shortstraw: A simple and effective corner finder for polylines. In: SBM. pp. 33–40 (2008)
35. Xiong, Y., LaViola Jr, J.J.: Revisiting shortstraw: improving corner finding in sketch-based interfaces. In: Proceedings of the 6th Eurographics Symposium on Sketch-Based Interfaces and Modeling. pp. 101–108 (2009)
36. Xu, P., Huang, Y., Yuan, T., Pang, K., Song, Y.Z., Xiang, T., Hospedales, T.M., Ma, Z., Guo, J.: Sketchmate: Deep hashing for million-scale human sketch retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8090–8098 (2018)
37. Zbyszynski, M., Donato, B.D., Tanaka, A.: Gesture-Timbre Space: Multidimensional Feature Mapping Using Machine Learning & Concatenative Synthesis p. 13 (2019)