ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

timbre.fun:
A gamified interactive system for crowdsourcing a timbre semantic vocabulary

Ben HAYES[1], Charalampos SAITIS[2], György FAZEKAS[3]

[1]Queen Mary University of London, United Kingdom, [b.j.hayes, c.saitis, gyorgy.fazeaks]@qmul.ac.uk

ABSTRACT

We present timbre.fun (https://timbre.fun/), a web-based gamified interactive system where users create sounds in response to semantic prompts (e.g., bright, rough) through exploring a two-dimensional control space that maps nonlinearly to the parameters of a simple hybrid wavetable and amplitude-modulation synthesizer. The current version features 25 semantic adjectives mined from a popular synthesis forum. As well as creating sounds, users can explore heatmaps generated from others' responses, and fit a classifier (k-nearest neighbors) in-browser. timbre.fun is based on recent work, including by the authors, which studied timbre semantic associations through prompted synthesis paradigms. The interactive is embedded in a digital exhibition on sensory variation and interaction (https://seeingmusic.app/) which debuted at the 2021 Edinburgh Science Festival, where it was visited by 197 users from 21 countries over 16 days. As it continues running online, a further 596 visitors from 35 countries have engaged. To date 579 sounds have been created and tagged, which will facilitate parallel research in timbre semantics and neural audio synthesis. Future work will include further gamifying the data collection pipeline, including "leveling-up" to unlock new words and synthesizers, and a full open-source release.

Keywords: Timbre, Synthesis, Crowdsourcing, Gamification, Timbre semantics

1 INTRODUCTION

The scientific study of musical timbre has historically focused primarily on understanding the perceptual phenomenon itself (28). This has variously involved seeking acoustic explanations for experimental data (13; 19; 33), studying its cross-modal associations (38; 15), and subjectively constructing systems for its description (25; 30). More recently, however, another approach has emerged which seeks to understand the relationship between timbre and sound production (15; 34) through *prompted synthesis* tasks. In these studies, participants are asked to produce a sound using a synthesiser in response to a prompt, such as a descriptive adjective. This yields a dataset of prompt-sound pairs which can be analysed to study the effects of various prompts on the sounds created.

This approach is of particular interest from the perspective of synthesiser control. The problem of providing meaningful controls for synthesisers has recurred in the literature since the late 70s (36; 27; 26), and has recently become of particular interest with the advent of complex neural audio synthesisers capable of being used in music production and sound design workflows (14; 21; 7; 2). However, the data collected in these studies have been limited in both scope and scale. Paired with the heterogeneity observed between timbral datasets (33), this has limited the usefulness of this data in such downstream tasks. To address these issues, we propose an approach for crowdsourcing timbre semantic data in a manner that is amenable to scaling and gamification. Through an exploratory analysis, we demonstrate that the data collected in a preliminary run of this study exhibit an emergent structure which is broadly congruent with the findings of prior timbre semantic research. We further observe, through the application of simple machine learning techniques, that the data collected in this manner hold sufficient predictive power to allow the affective connotations of semantic prompts to be classified from synthesiser parameters and acoustic features.

2 RELATED WORK

2.1 Prompted Synthesis

The standard paradigm for timbre semantic research involves listeners rating sounds along scales defined by descriptive adjectives (24). Controlling characteristics of the stimuli allows their perceptual influence on timbre semantic associations to be studied. This approach does not, however, provide insight into the inverse relationship: the influence of timbre perception and its semantic associations on the process of sound design. Despite significant effort having been invested into developing

adjective-controlled systems for these tasks (11; 12; 3; 32; 31; 9), this relationship has only recently begun to receive attention in the psychoacoustic literature (34; 15).

In the present authors' prior study (15), experienced sound designers programmed an FM synthesiser in response to adjectival prompts, and rated their created sounds on semantic scales. In Wallmark, et al.'s study (34), classically trained musicians interacted with a simplified 2D control space mapped to FM synthesiser parameters. In both studies, clear structure emerged when the created sounds were analysed in terms of their acoustic characteristics, synthesis parameters, and relationships to prompt words. Both similarities to and deviations from the findings of conventional timbre studies were observed, suggesting that the prompted synthesis approach is a viable method for studying the specifics of the "inverted" relationship between timbre and sound design.

2.2 Crowdsourcing

Estellés-Arolas and González-Ladrón-de-Guevara (10), through a detailed meta-analysis, defined crowdsourcing as "a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task." In this work, we view crowdsourcing as an alternative to lab-based or structured online methods of data collection, where increasing the flexibility of participation enables the collection of data at much greater scale.

Adjerid & Kelley (1) discuss how the allure of "big data" makes crowdsourcing attractive in experimental sciences, where logistical constraints often limit sample sizes. Where large scale datasets were once out of reach to all but the most highly resourced labs, they point out that a new generation of tools paired with the ubiquity of internet connections has brought this approach to data collection within the realm of possibility of more modestly financed projects. They argue that there are many benefits to collecting data at scale for multiple sub-fields of psychology, but go on to raise concerns around the ethics, quality, and appropriateness of these approaches to data collection.

Nonetheless, crowdsourced data has been shown to have benefits that extend beyond those induced simply by the size of the dataset. Casler, et al. (4) demonstrated that crowdsourcing a dataset resulted in greater socio-economic and ethnic diversity in the participant pool, whilst yielding almost indistinguishable test results from participants recruited in person or on social media. This suggests that crowdsourcing may be of particular interest in fields where cultural bias introduced through the homogeneity of participants is impactful. In timbre research, for example, it cannot be safely assumed that any perceptual phenomena observed are entirely devoid of cultural influence.

The similarity in group responses in Casler, et al.'s (4) study is encouraging in terms of the quality of data collected online. Nonetheless, the relince of psychoacoustic research on consistent audio reproduction sets the bar somewhat higher. To address this concern, Zacharakis, et al. (37) compared responses from a timbre dissimilarity study conducted in a laboratory, to those collected online. Timbre spaces computed from the two datasets showed very high configurational similarity, suggesting that both groups responded to similar acoustic cues despite variability in listening apparatus.

Crowdsourcing has previously been successfully applied to sourcing acoustic-semantic vocabularies in the context of music production. Specifically, Cartwright & Pardo (3) collected a set of descriptors of 40-band equalizer (EQ) curves from within an interactive EQ interface. Subsequently, Stables, et al. (32) captured descriptors from within the music production workflow by providing fully functional compressor, reverb, and EQ audio plugins.

2.3 Gamification

The term *gamification* broadly refers to the practice of imbuing a platform, application, or interface with motivational design features borrowed from games (20). It is of particular interest in the design of crowdsourced experiments, as these rely on a large body of participants being sufficiently motivated to provide responses. To address the ambiguity in terminology and praxis in crowdsourcing and gamification, Morschheuser, et al. (20) proposed a simple system of four archetypes for classifying gamified crowdsourcing systems, within which the present work would be described as a *crowd-creating* task.

Morschheuser, et al. state that a central challenge in designing a gamified crowdsourcing system is providing incentives that promote "the formation of positive motivations towards crowdsourcing work" whilst also fitting the type of activity. They further acknowledge that intrinsic motivation, such as that induced by tasks that "allow a participant to be creative and experience autonomy", can be dominated by the extrinsic motivation of social or financial reward, or perceived motivational affordances (e.g. credits or in-game rewards) of the task. For these reasons, we were careful in designing the present study to avoid rewards that might overshadow the central activity, or incentivise users to act quickly and carelessly. We took particular inspiration from the approach of Stables, et al. (32), who motivated users to interact with their audio plugins by providing access to the responses of other users. Further details of the choices we made are given in section 3.3.

With the growing number of online platforms for gamified crowdsourced data collection, the popularity of this approach is rapidly growing despite limited data as to its effectiveness (17). One study even found that gamification did not improve participant attrition rates for "effortful" and "unengaging" tasks (18). Further, Deterding, et al. (5) point out challenging,

unanswered ethical questions, including whether the "playful veneer" of games and gamified systems leads users to share more data than they otherwise would, and to what extent users' effort might be considered an unfair use of their labour. Despite these concerns, however, Keusch & Zhang (17) did observe a clear effect throughout the literature on psychological outcomes such as fun, interest, and satisfaction.

3 METHOD

3.1 Participants

Due to the casual, gamified nature of the experiment, participants were not subject to any selection or screening process, nor was any demographic information collected. Each participant was, however, assigned a unique anonymous user ID on loading the page which was stored with each response they submitted. On their first visit, users were asked if they would allow a cookie to be stored on their computer to allow this user ID to persist between sessions. If the cookie was declined, a new ID would be generated at each visit. At the time of writing, a total of 95 unique user IDs have been recorded, and a total of 579 responses submitted.

3.2 Web interface

The prompted synthesis interface was hosted online[1]. On arrival at the web page, visitors were given a brief explanation of the interface in informal language. Those who wished to learn more about the motivations behind the page were able to click a link to read some more detaihled information. After reading the introductory directions, participants were presented with a rectangular control space containing a black square. Clicking and dragging the black square around the space produced a sound from a synthesiser built on the WebAudio API. As the square moved around the space, the sound of the synthesiser changed in accordance with a control mapping. Above the control space was a short textual prompt, instructing participants to create a sound matching a given adjective. The created sound was stored in a database alongside a time series of points representing the path taken through the control space.

3.3 Gamification & rewards

As discussed in section 2.3, we opted for a "light touch" in terms of gamification and rewards. In particular, to create a more relaxed, game-like experience, we opted for a more colourful user interface with informal language, including the use of emoji, in the explanatory text. To maximise the intrinsic motivation of the task itself, we used language that emphasised its creative and exploratory nature, and designed the control space mapping to be sufficiently nonlinear as to prevent users from easily employing prior knowledge of sound synthesis. Finally, to provide a reward for creating sounds, two further features were unlocked once five trials had been completed: (i) explore the responses of other users via heatmap visualisations, (ii) fit a kNN classifier in-browser to predict the prompts associated with different sounds.

3.4 Word stimuli

The descriptive adjectives used in the textual prompts were taken from the ModWiggler corpus collected in a prior prompted synthesis study (15). As the sounds visitors created lacked a clear amplitude envelope or sub audio rate modulation and were thus effectively time invariant, we removed two words from the original list of 27 that clearly implied a particular temporal evolution: *percussive* and *plucky*. For a detailed explanation of how these were sourced, see (15).

3.5 Synthesiser

To encourage participants to explore the synthesiser control space by listening to the sounds rather than relying on prior expectations, we aimed to create a synthesiser and accompanying parameter mapping that fulfilled the following criteria:

1. sounds do not obviously resemble familiar instruments
2. control dimensions do not obviously map to familiar (e.g. FM or subtractive) synthesiser parameters
3. control dimensions do not ehibit an obvious linear mapping to any specific attribute of the sound

The resulting synthesiser design consisted of three components connected in series: an interpolated wavetable space, a ring modulator, and a bandpass biquad filter. Details of these components and the parameters they exposed are given below, and an overall schematic is given in Fig. 1.

To allow the synthesiser to produce a variety of harmonic distributions without implying a familiar synthesis method, a wavetable oscillator was chosen as the initial sound source, with wavetables sampled using trilinear interpolation from a 2-dimensional grid. This allows for a direct mapping from the 2D control space to the wavetable.

[1]The website is available at https://timbre.fun/

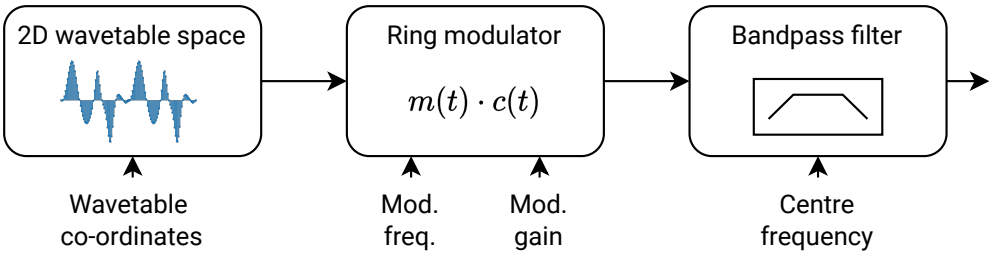

Figure 1. A schematic diagram of the hybrid wavetable and ring modulation synthesiser used in timbre.fun. The signal is generated by the wavetable oscillator, and processed by the ring modulator and bandpass filter. Parameter values are provided by nonlienar transformations of co-ordinates from the 2D control space.

This space of wavetables was sampled from a multilayer perceptron with sinusoidal activations (29) which was trained to reproduce single cycles of audio from the NSynth dataset (8). The network's inputs consisted of the x- and y-coordinates in the 2D wavetable space, and a time index within the wavetable, represented in two dimensions as a point on a unit circle to avoid discontinuities at the start and end of the wavetable. The x- and y-coordinate inputs for each training sample were derived as a learned matrix projection from a one-hot representation of sample identity, thus effectively allowing an unsupervised spatial organisation of wavetables to be learned. After training, the wavetable grid was sampled at a spatial resolution of 64 x 64 and a temporal resolution of 384.

The wavetable space allowed for the production of purely harmonic spectra, but was not capable of producing inharmonic signals. To address this, a ring modulator was added to the signal chain due to its ability to produce a finite number of inharmonic partials and its low dimensional parameter space. In a ring modulated signal, each frequency component ω_i in the carrier signal is replaced by two frequency components $\omega_i \pm \omega_m$ where ω_m is the modulator frequency. Ring modulation can therefore be used to create a variety of effects from subtle beating through to complete destruction of the harmonic structure of the signal. The modulation frequency and gain parameters were mapped to the 2D control space.

Prior research into the semantic correlates of timbral attributes (15; 38) suggests that a relationship exists between certain verbal descriptors and acoustic features describing the shape or distribution of energy in the frequency spectrum. A notable example is the relationship between descriptions of *brightness* and the *spectral centroid* (23), an audio feature describing the weighted mean of frequencies present in a signal. With the wavetable space and ring modulator alone, however, participants would ave been unable to shape the broader distribution of spectral energy beyond the pre-existing variation in harmonic amplitudes between the various wavetables. For this reason, we introduced a biquadratic bandpass filter with a fixed resonance of $\frac{\sqrt{2}}{2}$ to the end of the signal chain. Participants were thus able to more closely control the spectral distribution of their created sound. The centre frequency parameter was mapped to the 2D control space.

3.5.1 Control mapping

Previous prompted synthesis studies have allowed participants to interact both through full sets of synthesiser controls (15) and low dimensional mappings (34), in which synthesis parameters correspond directly to the *x*- or *y*-axis. These approaches risk encouraging participants, especially those with existing synthesis experience, to rely on their intuition about the workings of the synthesiser in formulating their responses. To ameliorate this effect, we opted to obscure the relationship between the dimensions of the control space and the parameters of the underlying synthesiser (with the exception of the *x*- and *y*-coordinates in the wavetable space) by using a nonlinear mapping of the form:

$$p_i = (0.5\cos(ax + by + cxy) + 0.5)^2$$

where the parameters a, b, and c were unique to each synthesiser parameter. These parameters were manually tuned to ensure a subjectively wide variety of sounds across the space, with no clear linear relationship between either axis and any synthesiser parameter. The values used in the experiment are listed in the source code repository[2] alongside the final ranges of the resulting parameters.

3.6 Experiment design

Prior timbral studies have typically followed a design that facilitates statistical hypothesis testing, or the application of other relevant data analysis techniques. In particular, stimuli are usually applied equivalently to all participants, or at least to

[2] https://github.com/ben-hayes/crossmodal-synthesis

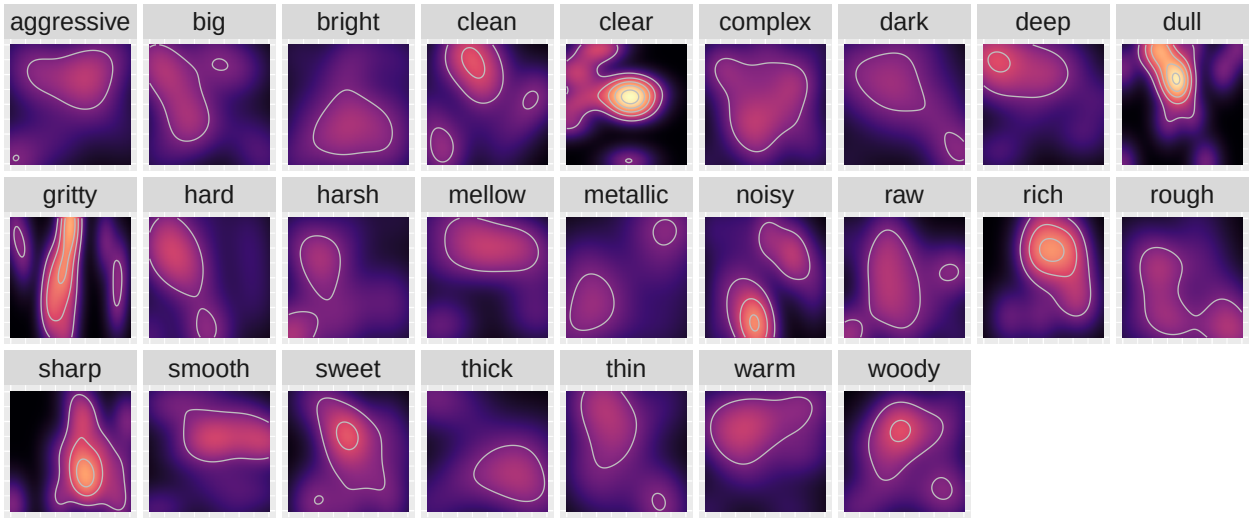

Figure 2. Heatmaps computed from 2D kernel density estimates of participants' responses in the control space per prompt.

those within a group. Whilst this is valuable from the perspective of drawing robust conclusions from the data, it constrains the scope of data that can be collected due to the redundancy that is necessitated by these statistical techniques. The utility of this data in applications that aim to use data-driven methods to build on research in timbre perception, such as synthesis control, is therefore limited.

As the purpose of *timbre.fun* is not to directly test hypotheses, but rather to pilot a method for crowdsourcing timbre data at scale with a view to supporting downstream applications, we opted to relax these usual constraints in favour of collecting a greater amount of more diverse data. Specifically, participants were able to complete as few or as many trials as they wished and prompts were simply sampled from a uniform distribution, with no guarantee that each participant responded to any specific subset of prompts.

4 RESULTS & DISCUSSION

Before analysis, the dataset was filtered to reduce the influence of low effort and accidental responses by removing any sound for which three or fewer JavaScript mouse move events were recorded. This resulted in a dataset containing 468 responses from 93 participants. Due to the inherently unmatched and unbalanced nature of the collected data, we forgo conventional statistical analysis in lieu of an exploratory analysis.

4.1 Control space density estimates

As an initial data exploration step, we visualised 2D kernel density estimates – with a standard bivariate normal – of participants' responses in the 2D control space, conditioned on the semantic prompt. These are shown in Fig. 2

For the majority of prompts, the resulting densities appear relatively diffuse and lack distinct modes, suggesting that participants found suitable responses across large regions of the control space. Certain prompts, however, do appear to have resulted in clearer agreement between participants: *clear*, *dull*, *noisy*, and *sharp*, for example, appear to have received a large proportion of their responses in small regions of the control space.

High level similarities, congruent with the groupings of descriptors observed in previous timbre semantic work (15; 38; 6), do also seem to emerge. For example, *dark*, *deep*, and *dull* appear to have most of their responses concentrated around a similar region, while *thin* and *thick* primarily occupy opposing regions of the space. The nonlinear mapping of synthesis parameters, however, means that any regions exhibiting similarity other than very distinct modes should be interpreted with caution, as neighbouring areas of the control space could result in dramatically different sounds.

4.2 Effect of prompt on synthesis parameters

Fig. 3 illustrates the distribution of responses for two of the five synthesiser parameters – filter centre frequency and modulator gain – conditional on the prompt given. In both cases the x-axis is sorted according to the median value per prompt, and the colour scale is normalised to the full available range of the parameter. Plots for the remaining prompts

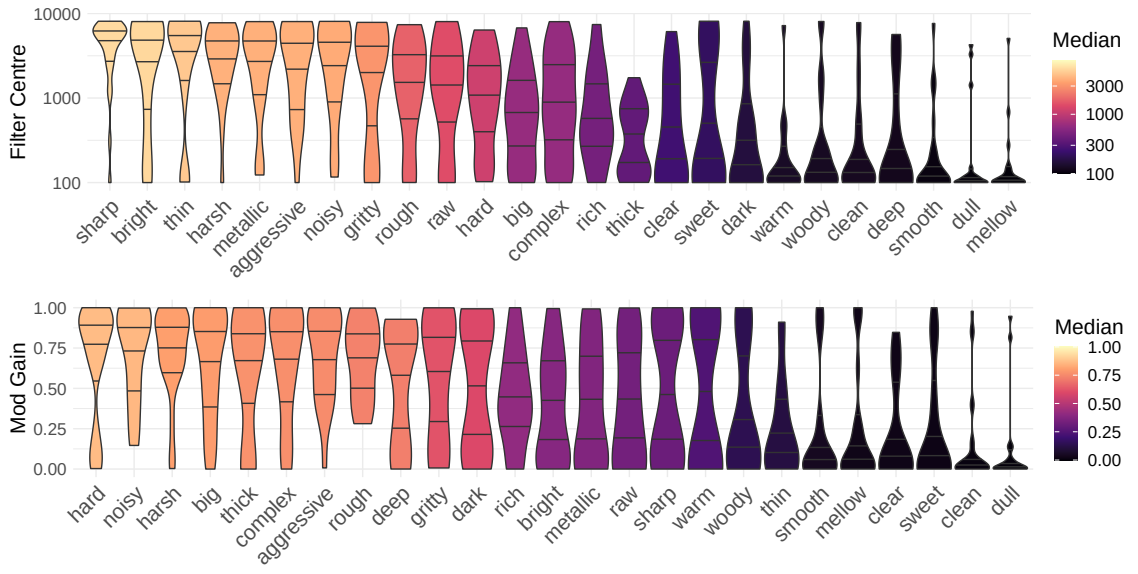

Figure 3. Effects of prompt on the filter centre frequency and modulator gain parameters.

are available online[3].

Very pronounced differences between prompts are visible for the Filter Centre and Modulation Gain parameters. In the case of Filter Centre, we see prompts including *sharp*, *bright*, *thin*, and *harsh* resulting in consistently high parameter values, and *deep*, *smooth*, *dull*, and *mellow* resulting in lower ones. For the Modulation Gain, we see *hard*, *noisy*, *harsh*, and *big* resulting in higher values, while *clear*, *sweet*, *clean*, and *dull* lead to lower values. These coarse groupings are, again, consistent with the factor structure of our prior prompted synthesis study, in which *sharp*, *bright*, and *harsh* all exhibited their highest loadings on the same factor, and *clean*, *clear*, and *sweet* all exhibited their highest on another factor. The grouping

Effects on the remaining parameters (see online supplement for plots) are markedly weaker, with wider distributions and less difference between median parameter values. Whilst the prompt effects on wavetable *x*- and *y*-coordinates appear primarily to have resulted in diffuse distributions, some prompts do appear to have a resulted in a concentration of responses around a particular value, suggesting that the wavetables at this point in the space may agree particularly well with the descriptor. These prompts include *sharp*, *gritty*, and *noisy*.

4.3 Acoustic features analysis

Whilst synthesiser parameters offer a comprehensive description of how to reproduce a given sound from the dataset, they provide only indirect insight into a sound's acoustic characteristics, owing to their perceptual nonlinearity and properties emerging from their interaction. We therefore extracted a set of acoustic features from all sounds created in the experiment to more completely describe them.

A four second long WAV file was rendered for each sound using the same WebAudio synthesiser as found on the `timbre.fun` website. Features were extracted from each audio file using the Timbre Toolbox (22). As the sounds lacked any temporal evolution, features computed on the Temporal Energy Envelope representation were excluded. All other representations and features included in the Toolbox's default configuration were used. By default, the Timbre Toolbox reports two summary statistics aggregated over time: the median and interquartile range. Due, again, to the lack of temporal variation in the sounds, we discared the interquartile range and retained only the median.

As many of the computed features measure very similar characteristics, we removed columns with high pairwise correlations in order to avoid the analysis being biased by over-representation of certain properties. In particular, for every pair of features with a Spearman correlation coefficient of greater than 0.9, we removed the feature with the higher mean absolute correlation across all remaining features. The list of remaining features is given in Table 1.

Principal components analysis was then applied to the remaining features. Parallel analysis (16) was used to select an appropriate number of components. This procedure involves repeating PCA on a number of randomly sampled, uncorrelated datasets of equivalent size, and retaining n components, where the n-th component in the real dataset is the component with the lowest eigenvalue that still exceeds a given percentile of the eigenvalues of the n-th components of the random

[3]See online supplement: `https://benhayes.net/timbre-fun/`

Table 1. Acoustic features remaining after removing collinear pairs and their loadings onto the first three principal components. Loadings with magnitude >0.7 are depicted in bold. See (22) for detailed descriptor explanations.

Rep.	Feature	PC1	PC2	PC3	Rep.	Feature	PC1	PC2	PC3
Harm.	Spec. Crest	**0.893**	0.009	0.011	Harm.	F0	0.176	–0.385	**–0.756**
Harm.	Spec. Skewness	**0.858**	–0.135	0.054	Harm.	Spec. Slope	–0.688	–0.269	–0.258
STFT	Spec. Skewness	**0.817**	–0.181	0.313	ERB	Spec. Crest	0.566	0.250	–0.399
Harm.	Spec. Decrease	**–0.762**	0.315	–0.224	Harm.	Inharmonicity	–0.488	0.331	0.571
ERB	Spec. Flatness	**–0.715**	–0.466	0.058	STFT	Spec. Flux	–0.371	0.317	0.356
Harm.	Harm. Spec. Dev.	0.254	**0.860**	–0.033	Audio	Auto-correlation	–0.345	0.243	–0.330
Harm.	Noise Erg.	0.457	**0.797**	–0.125	Harm.	Odd/Even Ratio	0.275	–0.132	0.087
ERB	Frame Erg.	–0.352	**0.730**	–0.151	Harm.	Spec. Flux	–0.089	0.249	–0.286

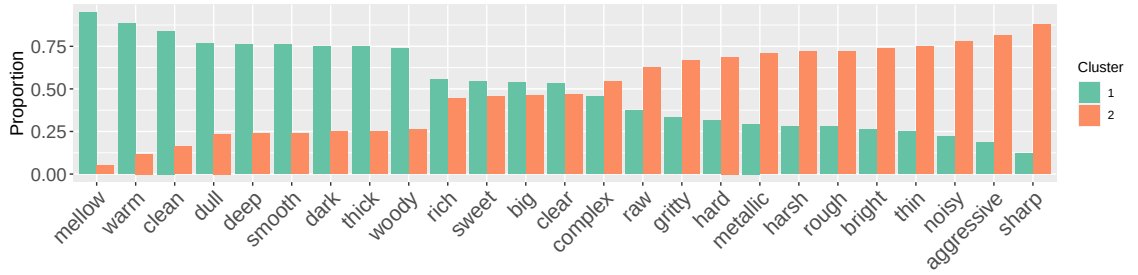

Figure 4. Assignments of prompt words to each of the two acoustic clusters. The leftmost (yellow) bar for each descriptor corresponds to cluster 1, and the rightmost (green) bar to cluster 2.

datasets. We ran the procedure for 480 iterations and found that three components exceeded the 99th percentile. Loadings of the first three components are reported in Table 1.

The resulting principal components were then subject to k-means clustering. A two-cluster solution was selected as optimal by computing the connectivity, Dunn index, and silhouette coefficient. For each of the 25 prompt words, we computed the proportion of sounds created in response to that word in each acoustic cluster. These are shown in Fig. 4. Note that the proportion was used in lieu of the raw count owing to the unbalanced design of the experiment. The majority of prompts appear to strongly associated with a specific cluster, with 19 prompts having at least twice as many responses in one cluster than the other.

We note that descriptors that have previously been associated with distinct semantic factors are grouped together in these clusters. For example, cluster 2 contains the majority of responses for *sharp*, *thin*, and *rough*, whilst these terms load strongly onto two distinct factors in our prior study (15) and onto three distinct factors in Zacharakis, et al.'s prior work (38). Instead of representing distinct timbral dimensions, the acoustic clusters instead appear to correspond to the opposite ends of an aggregation of timbral scales. This emergence of two timbral "poles" in this manner is somewhat consistent with the behaviour observed in our previous study (15) where, despite post-hoc semantic ratings supporting a five factor solution, correlations between semantic factors and synthesiser parameters appeared to form just two groups.

4.4 Word affect

Using the word affect norms collected by Warriner, et al. (35), valence, arousal, and dominance scores were obtained for all 25 prompts. In all cases except one the exact word form was found in the dataset. In the one case where this wasn't possible (woody), the closest adjectival form that shared a lemma (wooden) was used in its place. Each of the three dimensions was then discretised into two classes: *high* for those above the mean, and *low* for those below. Two classes were used instead of the three used by Wallmark, et al. (34) to enable both direct comparison with our acoustic clusters and the formulation of binary classification problems (see Section 4.5).

To examine the relationship between these affect classes and the clusters obtained from acoustic features, prompts were

Table 2. Test-set accuracy for all SVM models. Columns signify input features and rows signify targets. p-values taken from binomial test with no-information rate as null hypothesis. *: $p<0.05$, ***: $p<0.001$

	External Cluster Validation			*SVM Test Accuracy (%)*	
Dimension	Rand	Jaccard	Fowlkes-Mallows	Synth. Params	Acoustic PCs
Valence	0.667	0.485	0.653	61.3	62.4
Arousal	0.847	0.725	0.840	73.1***	71.0***
Dominance	0.667	0.485	0.653	53.8	62.4*

assigned to the acoustic cluster which contained a greater proportion of their responses, and three external validation metrics – the Rand, Jaccard, and Fowlkes-Mallows indices – were computed between these assignments and each of the three affect dimensions in turn. The results of these metrics are presented in Table 2. Whilst low to moderate values were observed for the valence and dominance classes, the arousal classes showed very high similarity to the separation of prompts between acoustic clusters. Whilst neither the data nor analysis are sufficient to establish a causal relationship, these results are strongly suggestive of an interaction between prompt arousal and the acoustic characteristics of the resulting sound.

4.5 Affect classification with support vector machines

As discussed previously, the structure of this dataset precludes conventional statistical hypothesis testing. Therefore, to further examine the relationship between the created sounds and the affective connotations of the prompt words we instead opted to fit a predictive machine learning model. In particular, we used a support-vector machine (SVM) with radial basis functions. We trained separate models on both acoustic principal components and the raw synthesiser parameters.

Data was partitioned into 80% training and 20% test subsets. Using this split, six models were fit — one for each combination of affect dimension and feature set (where synth parameters and acoustic principal components were the two feature sets). SVM hyperparameters were tuned using the adaptive random search implemented in the `caret` package for R, following a 10-fold cross validation with 5 repeats of each fold. To ensure an even class balance, training folds were upsampled by randomly repeating samples from the less represented class.

Table 2 lists the test-set accuracy of each trained SVM. Statistical significance of the accuracy was determined using a binomial test, where the no-information rate of the test dataset (i.e. the raw class proportions) was used as the null hypothesis. Consistently with the external validation metrics, the test accuracy observed for arousal classes was highly statistically significant, further supporting an affect-timbre interaction.

These results are distinct from those of Wallmark, et al. (34), who did not observe a significant effect of arousal on acoustic principal components. Deeper comparison of these results is challenging due to methodological differences, but such a clear deviation from the particular timbre-affect relation observed in their work is nonetheless noteworthy. In particular, it both lends further weight to the role of affect in prompted timbre production, whilst also suggesting that the mechanism of this interaction may to some extent be influenced by the method of synthesis available.

5 CONCLUSION

This paper presented a method for sourcing a timbre semantic vocabulary of prompt-sound pairs at scale using a gamified crowdsourced approach. A pilot run of the method, debuted as part of a digital exhibition on sensory variation and interaction at the 2021 Edinburgh Science Festival, resulted in a total of 579 sounds being created by 95 users. Exploratory analysis of the collected data suggests that responses showed logical consistency with prior timbre research, and exhibit sufficient structure to allow certain affective connotations of prompt words to be predicted from acoustic characteristics of the created sounds. These results support further work on gamified crowdsourced collection of timbral datasets, with a view to producing large datasets for use in downstream tasks, including descriptive synthesiser control.

In future work, we will augment the data collection pipeline by introducing new sound synthesis methods, new types of interaction (including tagging, dissimilarity rating, and more), and further gamification elements (such as a reward/points system, and the ability to unlock new features). We will also conduct a run of the study at a larger scale, and provide a full open-source release of the platform so that it can be adapted for other data collection tasks.

ACKNOWLEDGEMENTS

This work was supported by UK Research and Innovation [grant number EP/S022694/1].

REFERENCES

[1] I. Adjerid and K. Kelley. Big data in psychology: A framework for research advancement. *American Psychologist*, 73(7):899–917, Oct. 2018.

[2] A. Caillon and P. Esling. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv:2111.05011 [cs, eess]*, Dec. 2021. arXiv: 2111.05011.

[3] M. B. Cartwright and B. Pardo. Social-eq: Crowdsourcing an equalization descriptor map. In *Proceedings of the 14th international society for music information retrieval conference*, pages 395–400, Jan. 2013.

[4] K. Casler, L. Bickel, and E. Hackett. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6):2156–2160, Nov. 2013.

[5] S. Deterding, A. Canossa, C. Harteveld, S. Cooper, L. E. Nacke, and J. R. Whitson. Gamifying Research: Strategies, Opportunities, Challenges, Ethics. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2421–2424, Seoul Republic of Korea, Apr. 2015. ACM.

[6] T. M. Elliott, L. S. Hamilton, and F. E. Theunissen. Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones. *The Journal of the Acoustical Society of America*, 133(1):389–404, Jan. 2013.

[7] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts. DDSP: Differentiable Digital Signal Processing. In *8th International Conference on Learning Representations*, Addis Ababa, Ethiopia, Apr. 2020.

[8] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan. Neural audio synthesis of musical notes with WaveNet autoencoders. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 1068–1077, Sydney, Australia, Aug. 2017.

[9] P. Esling, N. Masuda, A. Bardet, R. Despres, and A. Chemla-Romeu-Santos. Flow Synthesizer: Universal Audio Synthesizer Control with Normalizing Flows. *Applied Sciences*, 10(1):302, Dec. 2020.

[10] E. Estellés-Arolas and F. González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2):189–200, Apr. 2012.

[11] R. Ethington and B. Punch. SeaWave: A System for Musical Timbre Description. *Computer Music Journal*, 18(1):30, 1994.

[12] A. Gounaropoulos and C. Johnson. Synthesising timbres and timbre-changes from Adjectives/Adverbs. In *Applications of evolutionary computing. EvoWorkshops 2006*, pages 664–675, Berlin, Heidelberg, Apr. 2006.

[13] J. M. Grey. Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5):1270–1277, 1977.

[14] B. Hayes, C. Saitis, and G. Fazekas. Neural Waveshaping Synthesis. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, Online, Nov. 2021.

[15] B. Hayes, C. Saitis, and G. Fazekas. Disembodied Timbres: A Study on Semantically Prompted FM Synthesis. *Journal of the Audio Engineering Society*, 70(5):373–391, May 2022.

[16] J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, June 1965.

[17] F. Keusch and C. Zhang. A Review of Issues in Gamified Surveys. *Social Science Computer Review*, 35(2):147–166, Apr. 2017.

[18] J. Lumsden, A. Skinner, D. Coyle, N. Lawrence, and M. Munafo. Attrition from Web-Based Cognitive Testing: A Repeated Measures Comparison of Gamification Techniques. *Journal of Medical Internet Research*, 19(11):e395, Nov. 2017.

[19] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3):177–192, Dec. 1995.

[20] B. Morschheuser, J. Hamari, J. Koivisto, and A. Maedche. Gamified crowdsourcing: Conceptualization, literature review, and future agenda. *International Journal of Human-Computer Studies*, 106:26–43, Oct. 2017.

[21] J. Nistal, C. Aouameur, I. Velarde, and S. Lattner. DrumGAN VST: A Plugin for Drum Sound Analysis/Synthesis With Autoencoding Generative Adversarial Networks, June 2022. arXiv:2206.14723 [cs, eess].

[22] G. Peeters, B. L. Giordano, P. Susini, N. Misdariis, and S. McAdams. The Timbre Toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, 130(5):2902–2916, Nov. 2011. Publisher: Acoustical Society of America.

[23] C. Saitis and K. Siedenburg. Brightness perception for musical instrument sounds: Relation to timbre dissimilarity and source-cause categories. *The Journal of the Acoustical Society of America*, 148(4):2256–2266, Oct. 2020.

[24] C. Saitis and S. Weinzierl. The Semantics of Timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay, editors, *Timbre: Acoustics, Perception, and Cognition*, volume 69, pages 119–149. Springer International Publishing, Cham, May 2019. doi: 10.1007/978-3-030-14832-4_5.

[25] P. Schaeffer, C. North, and J. Dack. *Treatise on musical objects: essays across disciplines*. Number 20 in California studies in 20th-century music. University of California Press, Oakland, California, 2017.

[26] A. Seago. A New Interaction Strategy for Musical Timbre Design. In S. Holland, K. Wilkie, P. Mulholland, and A. Seago, editors, *Music and Human-Computer Interaction*, pages 153–169. Springer London, London, 2013. Series Title: Springer Series on Cultural Computing.

[27] A. Seago, S. Holland, and P. Mulholland. A Critical Analysis of Synthesizer User Interfaces for Timbre. In *Proceedings of the XVIII British HCI Group Annual Conference HCI 2004*, volume 2, pages 105–108, Bristol, UK, Sept. 2004. Research Press International.

[28] K. Siedenburg, C. Saitis, and S. McAdams. The Present, Past, and Future of Timbre Research. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay, editors, *Timbre: Acoustics, Perception, and Cognition*, pages 1–19. Springer International Publishing, Cham, May 2019.

[29] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein. Implicit Neural Representations with Periodic Activation Functions. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., June 2020.

[30] D. Smalley. Spectromorphology: explaining sound-shapes. *Organised Sound*, 2(2):107–126, Aug. 1997.

[31] R. Stables, B. De Man, S. Enderby, J. Reiss, T. Wilmering, and G. Fazekas. Semantic description of timbral transformations in music production. In *ACM multimedia, oct. 15-19, amsterdam, netherlands*, pages 337–341, Oct. 2016.

[32] R. Stables, S. Enderby, B. De Man, G. Fazekas, and J. D. Reiss. SAFE: A system for extraction and retrieval of semantic audio descriptors. In *Proceedings of the 15th international society for music information retrieval conference*, Taipei, Taiwan, Oct. 2014.

[33] E. Thoret, B. Caramiaux, P. Depalle, and S. McAdams. Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre. *Nature Human Behaviour*, Nov. 2020.

[34] Z. Wallmark, R. J. Frank, and L. Nghiem. Creating novel tones from adjectives: An exploratory study using FM synthesis. *Psychomusicology: Music, Mind, and Brain*, 29(4):188–199, July 2019.

[35] A. B. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, Feb. 2013.

[36] D. L. Wessel. Timbre Space as a Musical Control Structure. *Computer Music Journal*, 3(2):45, June 1979.

[37] A. Zacharakis, B. Hayes, C. Saitis, and K. Pastiadis. Evidence for timbre space robustness to an uncontrolled online stimulus presentation. In *Proceedings of the 2nd International Conference on Timbre*, Thessaloniki, Greece (Online), Sept. 2020.

[38] A. Zacharakis, K. Pastiadis, and J. D. Reiss. An Interlanguage Study of Musical Timbre Semantic Dimensions and Their Acoustic Correlates. *Music Perception: An Interdisciplinary Journal*, 31(4):339–358, Apr. 2014.