OPEN ACCESS Freely available online

B. Hayes, C. Saitis, and G. Fazekas, "Disembodied Timbres: A Study on Semantically Prompted FM Synthesis"
J. Audio Eng. Soc., vol. 70, no. 5, pp. 373–391, (2022 May).
DOI: https://doi.org/10.17743/jaes.2022.0006

Disembodied Timbres: A Study on Semantically Prompted FM Synthesis

BEN HAYES, CHARALAMPOS SAITIS, AND GYÖRGY FAZEKAS, *AES Associate Member*
(b.j.hayes@qmul.ac.uk) (c.saitis@qmul.ac.uk) (g.fazekas@qmul.ac.uk)

Centre for Digital Music, Queen Mary University of London, United Kingdom

Disembodied electronic sounds constitute a large part of the modern auditory lexicon, but research into timbre perception has focused mostly on the tones of conventional acoustic musical instruments. It is unclear whether insights from these studies generalize to electronic sounds, nor is it obvious how these relate to the creation of such sounds. This work presents an experiment on the semantic associations of sounds produced by FM synthesis with the aim of identifying whether existing models of timbre semantics are appropriate for such sounds. A novel experimental paradigm, in which experienced sound designers responded to semantic prompts by programming a synthesizer, was applied, and semantic ratings on the sounds they created were provided. Exploratory factor analysis revealed a five-dimensional semantic space. The first two factors mapped well to the concepts of luminance, texture, and mass. The remaining three factors did not have clear parallels, but correlation analysis with acoustic descriptors suggested an acoustical relationship to luminance and texture. The results suggest that further inquiry into the timbres of disembodied electronic sounds, their synthesis, and their semantic associations would be worthwhile and that this could benefit research into auditory perception and cognition and synthesis control and audio engineering.

0 INTRODUCTION

The term "timbre" refers to a set of perceptual attributes that listeners use to discriminate different sounds, in addition to pitch, loudness, duration, spatial position, and the acoustic environment. Timbre is an inescapable component of the auditory experience. It enables listeners to identify who is speaking to them and ascertain the source of a sound, and it is of central importance to the aesthetic experience of music [1].

Increasingly the timbral world is populated by sounds with no discernible physical source, which are referred to in this article as *disembodied sounds*. Contemporary sound design tools and sound reproduction apparatus pair to enable listeners to experience sounds seemingly unconstrained by the acoustics of a physically resonating body. Such sounds now permeate day-to-day life in the form of notifications and alerts, heighten the visceral satisfaction received from movies and games, and have defined entirely new audio cultures [2]. Scientific understanding of timbre, however, is largely limited to insights gleaned from studies on musical instrument sounds playing isolated notes. This work sets out to systematically examine sounds that lack the kind of source-cause associations afforded by musical instruments through a novel experimental paradigm in which participants synthesize electronic sounds prompted with well-established semantic dimensions of timbre.

Studying the perception of disembodied electronic sounds may help further elucidate the mechanisms underpinning the experience of timbre [3, 4]. Specifically the way such timbres are talked about can disclose significant information about the way they are perceived [5, 6]. Common semantic dimensions for musical instrument sounds have been summarized as brightness/sharpness (or *luminance*), roughness/harshness (or *texture*), and fullness/richness (or *mass*) [7]. A primary aim of this study was to ascertain whether such dimensions are sufficient to describe the timbral variability of sounds produced by an FM synthesizer. This study also sets out to identify whether prompting synthesis with semantic descriptors would result in a discernable impact on the control of synthesizer parameters.

Beyond psychoacoustic insight, inquiry into the perception of disembodied timbres can inform further research in audio engineering and sound design. Many of today's

most popular software and hardware synthesizers do not represent a significant progression from the approach of early synthesizers—their controls continue to direct the synthesis at a low level, with complex systems of interdependence, limiting the ability of musicians and sound designers to predictably alter the perceptual attributes of a sound [8]. Previous work aiming to facilitate synthesis control by mapping from a conceptual representation, such as a timbre dissimilarity space [9, 10], high level features [11], or spatial representations of source-cause cues [12, 13] has focused on perceptual insights from research on acoustic sound sources. Thus studying the perception of disembodied timbres may also lead to insights into how synthesis control can be improved to more closely map to perception. To facilitate further research in this direction, the dataset of sounds generated in this study has been made available.[1] Full parameter configurations, semantic ratings, acoustic features, and anonymous participant questionnaire responses are provided, alongside rendered audio of all synthesized sounds.

0.1 From Sounds to Adjectives

The perception of timbre has enjoyed an extensive lineage of scientific inquiry, dating at least as far back as Helmholtz's [14] treatise *On The Sensations of Tone*. It is widely agreed to be a multi-faceted percept, and so two prevailing approaches to its study—perceptual and semantic—both seek dimensional decompositions of the timbre gestalt [1].

The first approach aims to directly tap into the perceptual structure of timbre by collecting pairwise general dissimilarity ratings on a set of sounds. Multidimensional scaling (MDS) techniques are then applied to recover a spatial configuration known as "timbre space" in which the distance between points corresponds to their perceived timbral difference. Today a number of MDS studies have confirmed at least two robust perceptual dimensions of timbre [15–18]. These correlate well with the duration of the attack part of the temporal envelope and the center of gravity of the spectral envelope, respectively. Additional dimensions appear to depend on the specific stimulus set. More recently a study applied a biologically inspired model that involved learning kernel distance functions over data from 17 previous dissimilarity studies [19]. Results showed that as well as sharing general acoustic correlates, each study's dataset yielded a number of experiment-specific correlates, suggesting that care should be taken in generalizing the results of any particular timbre study.

The second approach involves studying timbre perception indirectly through its semantic associations, that is, how language is employed to describe the timbre of a sound via cross-modal, onomatopoeic, or abstract metaphor [7]. Building on the underlying assumption that the perceptual attributes of timbre are encapsulated in its verbal descriptions, dimensionality reduction techniques, such as exploratory factor analysis and principal components analysis (PCA), are used to construct semantic timbre spaces from ratings of stimuli along verbally anchored scales. These are typically constructed either by two opposing descriptive adjectives, such as "rough-smooth" (known as the semantic differential method [20]), or an adjective and its negation, as in "rough–not rough" (known as the verbal attribute magnitude estimation method [21]).

This approach has a long history in empirical research on timbre, being first used in 1958 to study sonar sounds [22], about a decade before the early MDS studies of the 1970s [15, 16]. It was first applied to musical sounds in 1974 by von Bismarck [23], who used synthetic recreations of instrumental and vocal timbres. It has since been employed in numerous studies of musical timbre [24, 21, 25, 18, 26] (for a comprehensive review, see [7]). Despite differences in methodology (choice of verbal scales and dimensionality reduction technique) and stimuli, there is clear similarity between the semantic dimensions recovered by many of these studies. Typically a low-dimensional semantic space of timbre can be interpreted in terms of brightness/sharpness, roughness/harshness, and fullness/richness, although the precise demarcations between dimensions vary [7].

Zacharakis et al. [26] performed an interlanguage study with musically experienced Greek and English-speaking listeners, where responses from both linguistic groups were well explained by a model that also exhibited these three semantic dimensions. It was named the *luminance-texture-mass* (LTM) model based on the strongest factor loadings from both languages. A confirmatory study [27] using two representative scales (highly loaded) for each of the three factors, conducted with the same stimuli but Greek listeners only, suggested the model was broadly effective for predicting both semantic ratings and pairwise dissimilarities. However the attack-time dimension emerging from analysis of pairwise dissimilarities, which differentiates more impulsive from more sustained temporal envelopes, could not be directly captured by the LTM dimensions.

More recently a 20-dimensional model has been proposed, derived from a mixture of interviews with and semantic ratings by professional orchestral musicians, including conductors and composers [28]. They were asked to imagine orchestral instrument sounds rather than listen to recorded stimuli, which allowed tapping into richer and more creative linguistic descriptions. The model dimensions include *rumbling/low/thick* (L/M), *soft/singing* (T), *watery/fluid*, *direct/loud*, *nasal/reedy* (M), *shrill/harsh/noisy* (L/T), *percussive* (P), *pure/clear*, *brassy/metallic* (L/T), *raspy/grainy* (T), *ringing/long decay*, *sparkling/brilliant* (L), *airy/breathy*, *resonant/vibrant*, *hollow* (M), *woody*, *muted/veiled*, *sustained/even* (P), *open*, and *focused/compact*. The parenthetical initials indicate potential correspondence to the three LTM factors [26]; "P" indicates dimensions that relate to contrasting temporal envelope types (percussive and sustained).

The majority of this research focuses on physical instruments from the Western tonal music canon. When electronic and synthesized sounds do find use, it is typically either for the purposes of simulating the sounds of familiar

[1]The semantic FM dataset is available on *Zenodo*: https://doi.org/10.5281/zenodo.4609790.

acoustic instruments and the human voice [23, 24] or for the creation of controlled stimuli designed to elicit a specific perceptual response [29, 30]. It is not currently clear how well these multidimensional semantic models might generalize to more abstract and disembodied sounds, of the kind that increasingly populate the audio cultures of today. To this end, a study of electronic and electroacoustic "textural" sounds indicated a five-dimensional semantic space: *ordered–chaotic*, *homogeneous–heterogeneous*, *tonal–noisy*, *high/bright–low/dull* (L), and *smooth–coarser* (T) [5]. Two of these dimensions suggest that luminance and texture might generalize beyond the musical instrument domain. However the tested textural sounds involved multiple different timbres, iterative envelopes, and/or varying pitch profiles, all of which may not be suitable to examine the intrinsic dimensions of timbre per se, as indeed attested by the labels of the other three dimensions.

0.2 From Adjectives to Sounds

In the research discussed so far, the standard paradigm involves listeners rating a set of sounds along scales defined by descriptive adjectives. Stimuli are manipulated along one or more acoustical dimensions, and the aim is to explain their perceptual effect on semantic associations. However this method does not address the relationship between timbre and language from the opposite direction: How does the perceptual experience of timbre, through its semantic associations, relate to the creative process of sound design and engineering? In other words, how do semantic associations modulate acoustical response? This important question has received considerably less attention in the psychoacoustical literature, despite many relevant efforts to develop intuitive, adjective-controlled interfaces for audio synthesis and production [31–36]. To explore this question, here a semantically prompted FM synthesis task was used, and semantic associations of timbre were examined through their acoustical imprints on the creation of new sounds, effectively reverse engineering the standard paradigm.

Controlling the generation of complex audio spectra was made significantly easier by the invention of FM synthesis. Introduced by Chowning [37] in 1973, it generates rich spectra with nuanced patterns of spectral energy distribution. Strictly speaking, FM synthesis as formulated by Chowning, and as subsequently implemented in numerous commercial synthesizers, applies phase modulation rather than frequency modulation. That is, the carrier sinusoid is modulated by way of an additive term, rather than a multiplicative one. Pairing each oscillator with an amplitude envelope allows for further control of the spectrotemporal evolution of a sound. An FM synthesizer can be highly timbrally expressive with only a small number of oscillators and thus a limited number of parameters.

FM synthesis quickly found application in a variety of commercial synthesizers, including Yamaha's legendary DX7, and its timbral palette became highly influential on popular music over the subsequent decades but also in timbre research. In their 1995 timbre dissimilarity study, McAdams et al. [17] used simulations of traditional Western instruments synthesized by Wessel et al. [38] on a Yamaha TX802 FM Tone Generator. An earlier study of timbre semantics by Ashley [39] involved an FM system that "learned" to map certain controls with adjectives from users' verbal descriptions to changes in timbre.

FM timbres, therefore, are ideal as an object of study. They can be familiar enough as sonic entities to be distinctly identifiable and attract a varied aesthetic vocabulary while being abstract enough to avoid inherently implying a distinct source cause. Wallmark et al. [40] were the first to task a sample of classically trained musicians with creating a new timbre in response to adjectives sourced from orchestration books. To do so, participants explored a 2D space that linearly mapped to the controls of a simple FM synthesizer consisting of one modulator and one carrier. The experimental interface played a continuous tone at a fixed carrier frequency, whose spectral properties were shaped by the 2D controller. It also included a slider that controlled a distortion amplifier. Results suggested a relationship between word affect (valence and arousal) and certain distinct acoustical profiles. For instance, in response to both positive and negative high-arousal words such as brilliant or bright and rough or harsh, musicians crafted sounds with more strength in higher frequencies and inharmonicity.

0.3 The Present Study

The present study investigated how semantic associations modulate timbre perception (from adjectives to sounds) and vice versa (from sounds to adjectives) in the context of disembodied electronic sounds. These questions were approached by adapting the prompted synthesis paradigm [40] to enable comparative prompts (e.g., create a sound that is *rougher* or *less rough* than a played reference) followed by comparative ratings (e.g., rate how *much rougher* or *less rough* the created sound is from the reference). To promote ecological validity, adjectives were collected from an online message board for modular synthesizer enthusiasts, and the study focused on timbres created by music and audio technologists with experience in sound design and synthesis. Exploratory factor analysis of comparative semantic ratings and principal components analysis of acoustic features extracted from the created sounds were carried out. Linear regression and correlation analyses subsequently enabled quantification of the interrelations between language, psychoacoustics, and the adjustment of synthesizer controls.

Whereas the design of Wallmark et al. [40] focused solely on the effects of spectral energy distribution, because participants were shaping only static aspects of a continuous tone, the design of this study seeks to incorporate spectrotemporal and purely temporal aspects of the FM sounds by providing a full set of amplitude envelope controls. Three distinct fundamental frequency (F0) conditions for each comparative prompt were also applied. In research on timbre it is usual to equalize the F0 of stimuli as pitch and timbre are known to interact [41, 42]. In this study the authors wanted to explore whether such an interaction would exert an effect on synthesizer parameter control, that is, on shaping timbre.

The authors also wanted to examine the influence of F0 on the semantic dimensions of FM sounds.

1 METHOD

1.1 Participants

Thirty people took part in the experiment (mean age: $\mu = 28.7$ years; standard deviation: $\sigma = 7.52$ years; range: 21–55 years). All spent their formative years in an English-speaking country and self-reported prior synthesis experience via music production or sound design. They completed the Perceptual Abilities and Musical Training subscales of the Goldsmiths Musical Sophistication Index inventory [43]. Compared to the reference statistics provided with Goldsmiths Musical Sophistication Index, participants scored higher on Musical Training (this study: $\mu = 35.4$; reference study: $\mu = 26.5$) with a narrower distribution of scores (this study: $\sigma = 6.67$; reference study: $\sigma = 11.4$). Scores for Perceptual Abilities were slightly higher (this study: $\mu = 53.4$, $\sigma = 5.16$; reference study: $\mu = 50.2$, $\sigma = 7.86$). Participants gave written informed consent prior to the experiment. The study was approved by the Queen Mary Ethics of Research Committee (ref: QMREC2352a) and conducted in accordance with the Declaration of Helsinki.

1.2 Word Stimuli

To maximize the appropriateness of word stimuli selection to synthesized sounds, a corpus-based approach was adopted, which involved mining descriptors from a popular modular synthesis forum.[2] Publicly available posts from the forum dating up to February 21, 2020, were collected, for a total of 1,407,604 posts. After lemmatization, the corpus contained 330,700 unique tokens. Posts were filtered to a frequency-sorted list of words co-occurring in bigrams with the terms *sound*, *sounding*, *tone*, and *timbre*, which were then further filtered to retain only adjectives using Natural Language Toolkit's part of speech tagger. This resulted in a list of 96,277 potential descriptions of timbre, of which 5,977 were unique tokens. The 50 most frequently used timbral adjectives are displayed in APPENDIX A.1.

The list was independently pruned by two raters according to a set of criteria (given in APPENDIX A.2), resulting in a final set of 27 adjectives (see Table 1). To ensure variance along the LTM semantic dimensions, three descriptions were selected as prompts for the synthesis task, namely *bright*, *thick*, and *rough*. These were selected by filtering the set of 27 adjectives to only those that showed high loadings onto the English LTM factors in [26]. For example, *brilliant* and *bright* loaded highly onto the *luminance* factor. The word with the highest frequency in the corpus for each factor was then retained—e.g., *bright* in the case of the luminance factor.

1.3 Synthesizer

In its simplest form, FM synthesis can generate rich and complex timbres by time-varying the phase of an oscillator (carrier) via the output of a second oscillator (modulator) [37]. This is illustrated by Eq. (1):

$$x(t) = A \sin(\omega_c t + I \sin \omega_m t), \tag{1}$$

where A is the overall amplitude, ω_c the carrier frequency, ω_m the modulation frequency, and I the modulation index. Note that Eq. (1) strictly describes *phase modulation* rather than frequency modulation, which produces an equivalent magnitude spectrum when using sinusoidal oscillators. Because FM synthesizers are typically implemented with phase modulation, this formulation was used for the experimental synthesizer.

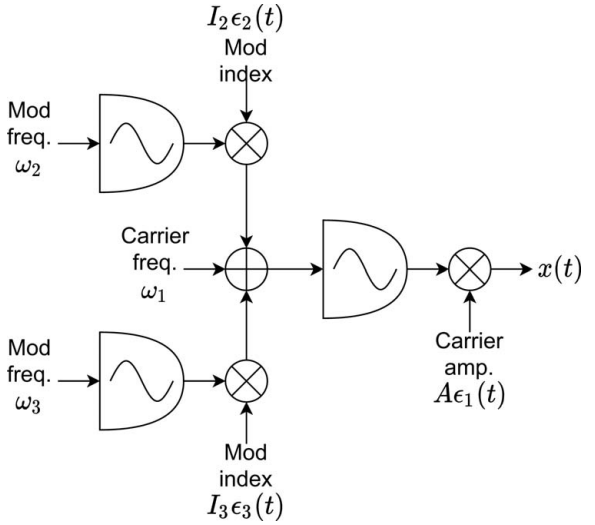

Fig. 1. A schematic diagram of the three-operator frequency-modulation synthesizer. Carrier amp., carrier amplitude; Carrier freq., carrier frequency; Mod freq., modulator frequency; Mod index, modulation index.

The synthesizer used in the experiment consisted of three sinusoidal oscillators (hereafter also referred to as operators) with an accompanying amplitude envelope and frequency modulation input. Operators 2 and 3 modulated the phase of operator 1 in linear combination (see Fig. 1). Each operator's amplitude was modulated by an independent Attack, Decay, Sustain, Release (ADSR) envelope. The attack portion was a linear ramp. The decay and release portions were exponential ramps where the segment length described the time taken to fall $1 - \frac{1}{e}$ of the way to the target value. The experimental synthesizer is thus given by Eq. (2):

$$x(t) = A\epsilon_1(t) \sin(\omega_1 t + I_2\epsilon_2(t) \sin \omega_2 t + I_3\epsilon_3(t) \sin \omega_3 t), \tag{2}$$

where ω_i gives the frequency of the ith operator, $\varepsilon_i(t)$ gives the amplitude envelope value of the ith operator at time t, and I_i gives the modulation index of the ith operator.

Participants were presented with a set of user controls for the FM synthesis parameters. In order to be consistent with the interfaces of popular FM synthesizers, the operator tuning ratio parameters were divided across two controls: *coarse* and *fine*. The *coarse* control specified the integer part of the tuning ratio, and the *fine* control specified the

[2] https://www.modwiggler.com/forum/.

Table 1. Factor loadings of semantic scales after Oblimin rotation. Suggested factor labels are given in parentheses.

	F1 (Sharpness)	F2 (Mass)	F3 (Clarity)	F4 (Percussiveness)	F5 (Rawness)
Sharp	**0.82**	–0.07	0.06	0.16	0.07
Metallic	**0.75**	0.05	–0.05	0.09	0.11
Bright	**0.73**	–0.22	0.04	0.10	0.05
Harsh	**0.72**	0.01	–0.18	0.08	0.15
Big	0.30	**0.87**	–0.03	–0.16	–0.04
Thick	–0.15	**0.84**	–0.10	0.02	–0.04
Deep	–0.43	**0.70**	–0.00	–0.07	0.06
Thin	0.32	**–0.70**	0.20	0.11	0.02
Clean	–0.04	0.02	**0.90**	–0.02	–0.01
Clear	0.17	–0.04	**0.78**	0.07	–0.03
Plucky	–0.04	–0.09	0.07	**0.99**	–0.05
Percussive	0.04	–0.02	–0.06	**0.78**	0.06
Raw	0.01	–0.12	0.12	0.01	**0.78**
Rich	0.32	0.69	0.08	–0.06	–0.03
Dull	–0.69	–0.12	0.02	–0.25	–0.03
Mellow	–0.67	–0.04	0.17	–0.12	–0.15
Woody	–0.63	0.20	0.01	0.23	–0.18
Warm	–0.60	0.42	0.17	–0.06	–0.01
Dark	–0.58	0.51	0.06	–0.05	0.19
Aggressive	0.57	0.27	–0.06	0.15	0.33
Sweet	–0.03	0.13	0.43	0.10	–0.56
Noisy	0.52	0.10	–0.40	0.11	0.12
Hard	0.49	0.24	–0.14	0.24	0.23
Smooth	–0.49	–0.00	0.40	–0.24	–0.08
Complex	0.48	0.36	–0.35	0.10	–0.11
Gritty	0.48	0.26	–0.32	0.18	0.17
Rough	0.42	0.16	–0.26	0.21	0.29

Note: Bold type indicates loadings with an absolute value greater than 0.70.

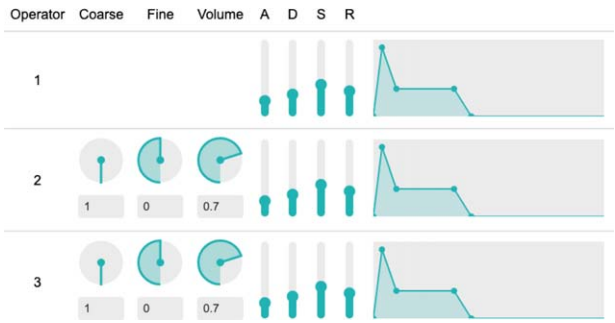

Fig. 2. The frequency-modulation synthesis interface used by the participants. Coarse, the integer component of the ratio of the modulator frequency to the carrier frequency; Fine, the fractional component (in 10^{-3} increments) of the ratio of the modulator frequency to the carrier frequency.

fractional part at a resolution of one thousandth. Dividing the controls in this way provides two benefits to the sound designer. Firstly they are able to quickly explore harmonic tuning ratios by fixing the *fine* control at zero. Secondly, because the sideband distribution is very sensitive to the tuning ratio, the precision of the *fine* control enables careful exploration of inharmonic values. In order to control for pitch and amplitude within trials, operator volume and tuning controls were only made available for modulating operators. This interface is shown in Fig. 2.

1.4 Procedure

Because of COVID-19, the study was conducted remotely. Participants accessed the experiment through a web browser and were instructed to use high quality headphones. Recent work suggests that timbre spaces constructed from pairwise dissimilarities collected online show good configurational similarity to those constructed from ratings collected in a laboratory setting [44]. The study was built using *lab.js* [45], and the WebAudio API's AudioWorklet was used to build a real-time in-browser FM synthesizer.[3]

The experiment consisted of a series of nine functionally identical trials, covering each combination of three comparative semantic prompts representing the LTM factors (*brighter* or *less bright*, *thicker* or *less thick*, and *rougher* or *less rough*) and three pitches (E2, A3, and D5) representing the low, middle, and high registers. The direction of comparison (less or more) was selected randomly each time (i.e., the number of trials was always nine). Each trial consisted of three steps:

1. A browser-based FM synthesizer was pre-set to generate a particular sound (the *reference* sound) with parameters p_r. Participants adjusted the controls to produce a new sound (the *created* sound) with pa-

[3]Source code for the study is available in a GitHub repository: https://github.com/ben-hayes/fm-synth-study.

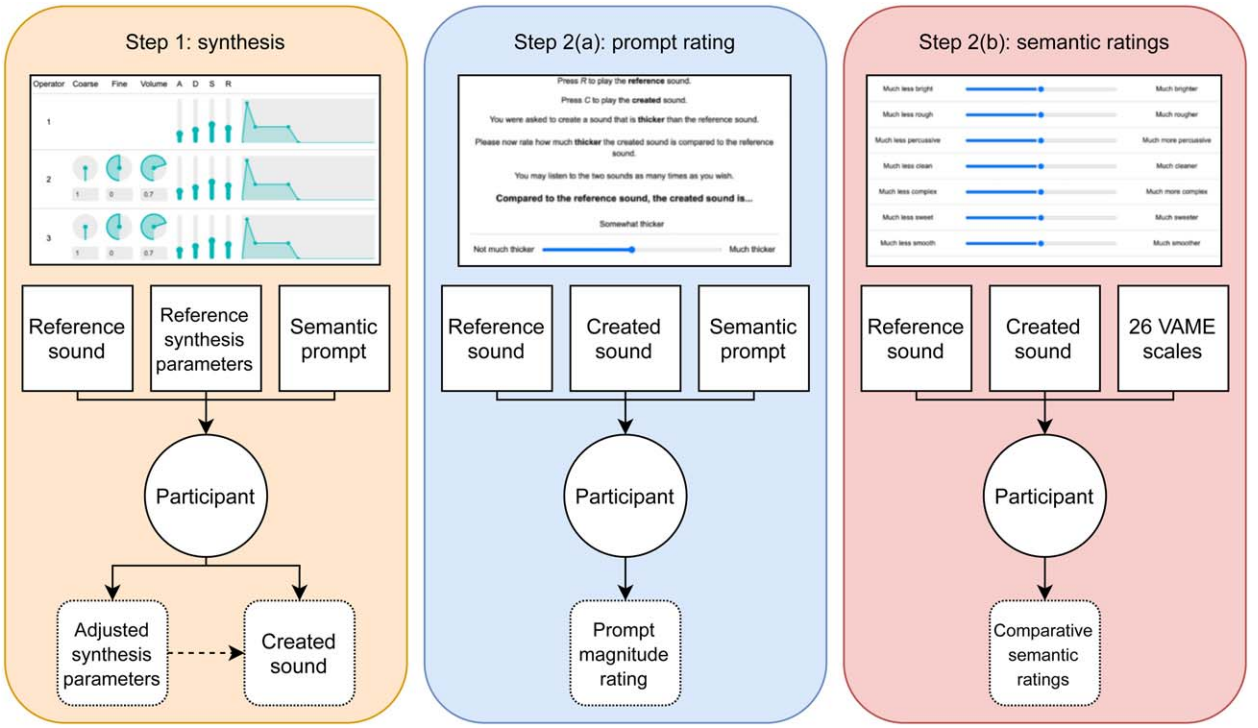

Fig. 3. A schematic diagram illustrating the experimental procedure for a single trial, repeated for each prompt and register. Step 1 (orange): participants synthesize a sound in response to a prompt. Step 2(a) (blue): participants rate the difference between the reference sound and their created sound in terms of the prompt. Step 2(b) (red): participants rate the difference between the reference sound and created sound in terms of 26 semantic descriptors.

rameters p_c to fulfil the given comparative prompt (e.g., to create a sound that is *brighter* or *less bright* than the reference).

2. Participants rated the magnitude of the difference between the sounds described by p_r and p_c in terms of the given prompt (e.g., how much *brighter* or *less bright* c is with respect to r). Ratings were input using a horizontal slider with a hidden range of 0.0 to 10.0 and resolution of 0.1.
3. Participants rated the magnitude of the difference between the sounds described by p_r and p_c in terms of the remaining two prompts (e.g., *thick* and *rough* if the initial prompt was *bright*) and the 24 additional timbral adjectives. Ratings were input using a horizontal slider with a hidden range of –10.0 to 10.0 and resolution of 0.1.

During each step, participants were able to listen to both the reference and created sounds as many times as they wished. There was no time limit imposed on any step. This procedure is illustrated in Fig. 3.

In each trial, the starting values of the synthesizer's parameters were given by randomly selecting an entry from the database of sounds created by previous participants. This approach enabled data to be collected on a wider range of parameter combinations than would be possible if the synthesizer was initialized identically for all participants. Given the sound design expertise of the participants, this approach also enabled the focus of the analysis to be on regions of synthesizer parameter space that are of interest to experienced synthesists. Limitations of this approach are discussed in SEC. 3.3. To start this process, the database was initialized with a starting set of nine "seed" sounds, which were hand-designed by the first author and loosely based on popular DX7 patches.

2 RESULTS

2.1 Exploratory Factor Analysis

Initial reliability analyses were conducted using Cronbach's α. All 27 semantic scales showed high internal consistency; average $\alpha = 0.95$ and $\sigma = 0.003$. Subsequently exploratory factor analysis was performed on the comparative ratings given across all 27 adjectives. Factor analysis is a technique for computing a set of latent factors from data, incorporating an independent stochastic error for each variable and observation. Each observation of a given variable can be considered as the sum of some amount of common variance (referred to as communality) and some amount of specific variance (consisting of any variance unique to that variable, plus any observation error).

To build a factor model from comparative ratings, it is assumed these are estimates of the difference between two unobserved absolute ratings $X_{\text{diff}} = X_c - X_r + \varepsilon_{\text{diff}}$, where

X_{diff} is the matrix of comparative ratings, X_c and X_r are matrices of the unobserved absolute ratings of created and reference sounds respectively, and $\varepsilon_{\text{diff}}$ is a normally distributed observation error of mean zero and finite variance. As a consequence of model linearity, it follows that a factor model of comparative ratings X_{diff} estimates the same loading matrix as a theoretical factor model of the unobserved absolute ratings given by a union of the elements of X_c and X_r (see APPENDIX A.3).

Selecting an appropriate number of factors is the subject of extensive discussion in the literature, and many methods remain in use. Fabrigar et al. [46] provide a review of such methods and a discussion of their strengths and weaknesses. Among the most popular are the Kaiser criterion, Cattell's scree test, and Horn's parallel analysis.

The Kaiser criterion [47] involves retaining as many factors as there are eigenvalues of the correlation matrix ≥ 1.0. In Cattell's [48] method, a scree plot (correlation matrix eigenvalues plotted against their indices) is inspected with the aim of identifying an "elbow" point that signifies an appropriate number of factors. Horn's parallel analysis [49] is a bootstrap method in which an identical factor analysis procedure is conducted on a large number of normally distributed random datasets of identical shape to the real data. The eigenvalues or sums of squared loadings (depending on the method) of the real data are then compared to a threshold statistic (usually the 95th percentile) from the randomly generated data. The number of values for which the real data exceeds the threshold statistic signifies the appropriate number of factors.

Empirical comparisons of these methods and others suggest that parallel analysis more reliably estimates the appropriate number of factors from both real [46] and synthetic [50] data. Conversely the Kaiser criterion consistently suggested a model with too few factors in the case of real data and too many factors when applied to synthetic data. With both real and synthetic data, the scree method was found to be variable in its accuracy and ambiguous in its interpretation. Accordingly here a semantic space for the created timbres was explored using parallel analysis, which supported a five-factor solution (Fig. 4). Factor analysis was performed using maximum likelihood estimation with non-orthogonal Oblimin rotation. A non-orthogonal rotation method was selected to avoid imposing assumptions about the independence of semantic factors. The factors cumulatively accounted for 74.36% of data variance. Individual factor variance is not available for the rotated solution because of the non-orthogonality of the factors.

The loadings of factors onto semantic descriptors are shown in Table 1. Factor F1 showed strong loadings onto terms associated with both luminance (including *sharp*) and texture (*metallic* and *harsh*). Factor F2 showed strong loadings onto terms related to mass (*big*, *thick*, and negatively *thin*). Factor F3 showed strong loadings for the words *clean* and *clear*, factor F4 for *plucky* and *percussive*, and factor F5 for *raw*. Proposed labels for each factor were chosen on the basis of either the highest-loading word (F1 and F5) or one that was deemed to better capture the meaning of the corresponding dimension (F2–F4).

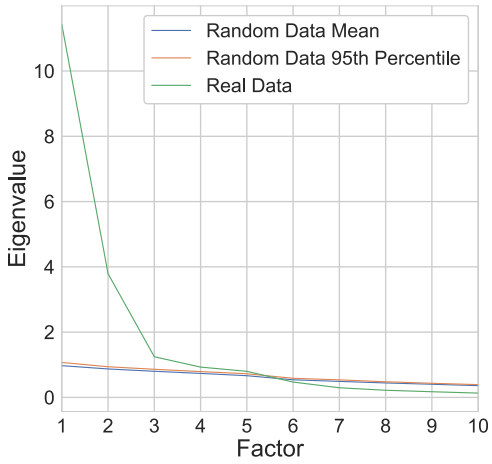

Fig. 4. A scree plot comparing the factor eigenvalues of the dataset to the mean and 95th percentile of the factor eigenvalues of the stochastic datasets generated in parallel analysis. Here it can be seen that the procedure supports five factors at the 95th percentile level.

Table 2. Inter-factor correlations and angles.

	F1	F2	F3	F4
F2	–0.08			
	94.4			
F3	–0.42	–0.30		
	114.6	*107.7*		
F4	0.51	–0.17	–0.27	
	59.3	*99.6*	*105.4*	
F5	0.37	0.07	–0.44	0.31
	68.3	*85.8*	*116.2*	*72.1*

Table 2 reports the inter-factor correlation coefficients (r) after rotation and the angles between rotated factors ($angle = cos^{-1}(r)$). There appeared to be moderate collinearity between F1 and F3–F5 and between F2 and F3, implying a degree of semantic entanglement across all five factors in the model. The lowest correlations were observed with F2, suggesting that impressions of *mass* in these FM sounds might have been perceptually more distinct from the other four semantic dimensions.

2.2 Acoustic Features Analysis

To study the psychoacoustic underpinnings of the semantic space, a large set of acoustic features were extracted from the created sounds (shown in Table 3). Spectral features were computed on multiple representations, namely short-time Fourier transform (STFT) magnitude and power spectra, Bark frequency magnitude spectrum, and harmonic peak magnitudes [51]. Furthermore harmonic features included inharmonicity, odd-to-even ratio, and tristimulus, and purely temporal features included log attack time, temporal centroid, and zero-crossing rate were computed.

Spectral features were computed using a Hamming window of size 1,024 with an overlap of 75%, and silent frames were discarded. Framewise features were summarized by the median and interquartile range. All features were computed using the Essentia library for Python. Synthesizer

Table 3. Extracted acoustic features.

Signal Representation	*Feature*	*Explanation*
$STFT_{mag}$ Spectrum	Centroid	Center of mass of spectral representation
$STFT_{pow}$ Spectrum	Spread	The statistical variance of the distribution of spectral energy
Bark Spectrum	Skewness	The asymmetry of the distribution of spectral energy
Harmonic Spectrum	Kurtosis	Proportional to the amount of energy in the tails of the spectral distribution
	Decrease	A linear regression coefficient representing the decreasing slope of the spectrum
	Rolloff	The frequency bin below which 85% of spectral energy is contained
	Frame Energy	The total energy contained in the spectrum
	Flatness	The ratio between the geometric and arithmetic means of the spectrum
	Crest	The ratio between the maximum value and arithmetic mean of the spectrum
Harmonic Peaks	Inharmonicity	The energy-weighted divergence of harmonic peak frequencies from integer multiples of the fundamental
	Tristimulus #1	Relative weight of first harmonic
	Tristimulus #2	Relative weight of second, third, and fourth harmonics
	Tristimulus #3	Relative weight of fifth harmonic and higher
	Odd-to-Even Ratio	Ratio of energy contained in harmonic peaks with odd index to energy in those with even index
	Noisiness	The difference between the total energy in the signal and the energy contained in harmonic peaks
Amplitude Envelope	Log Attack Time	The log (base 10) of the time taken for the signal to move from 20% to 90% of its maximum amplitude
	Effective Duration	The duration for which the signal is above 40% of its maximum amplitude
	Temporal Centroid	The center of mass of the amplitude envelope
Raw Waveform	Strong Decay	A nonlinear function of temporal centroid and signal energy
	Zero Crossing Rate	The proportion of signal values that represent sign changes

patches were rendered at 44.1 kHz with a duration of 4 s. The ADSR envelope was controlled by a gate signal, which was on (attack, decay, and sustain) for 3 s and off (release) for 1 s.

The extracted features cannot be assumed to correspond to independent axes of variation in the sounds under analysis. Indeed many features exhibit strong correlation. In order to address this issue, a feature dimensionality reduction procedure based on that of Zacharakis et al. [26] was followed. Their approach involved three reduction steps: Firstly they eliminated multicollinear features by inspecting Spearman rank correlation coefficients and discarding one member of any pair where $|\rho| > 0.8$. Secondly they inspected the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, defined as:

$$\mathrm{KMO}_i = \frac{\sum_{j \neq i} r_{ij}^2}{\sum_{j \neq i} r_{ij}^2 + \sum_{j \neq i} u_{ij}}$$

where R is the data correlation matrix and U is the data partial correlation matrix, that is, the correlations between pairs of variables controlling for the influence of other variables in the analysis. Variables with KMO <0.5 were discarded. Finally they performed PCA with Varimax rotation on the remaining features.

While this three-step method addresses the issue of correlated feature clusters, the remaining variables and, therefore, the structure of the resulting component space are highly dependent on which member of each collinear pair is retained in the first step. On several runs of the procedure with different orderings of variables in the first step, drastically different PCA solutions were found. Therefore, to improve reproducibility and select the most representative principal components, an extra step was introduced before the reduction procedure wherein features were sorted by their maximum absolute Spearman rank correlation coefficient with any of the semantic factors. Then the member of each collinear feature pair with the lowest such factor correlation was discarded. The authors believe this filter-based approach to be sufficient for the task of identifying acoustical correlates and thus leave deeper analysis of features and application of alternate feature selection methods to future work.

Because of the large number of features computed, the threshold for the Spearman rank correlation coefficient was set at 0.7 and for the KMO measure of sampling adequacy at 0.7. This resulted in a set of 17 descriptors, which are listed in Table 4. Parallel analysis, performed on the resulting set of features, supported a four-component solution at the 95th percentile level. PCA was followed by Varimax rotation to achieve simple structure. The resulting compo-

Table 4. Principal component (PC) loadings of acoustic features after varimax rotation.

	PC1 Spectrotemporal (Distribution) & Spectral Shape	PC2 Temporal Energy Variation & Spectral Slope	PC3 Spectrotemporal (Flatness)	PC4 Spectrotemporal (Crest Factor)
$STFT_{pow}$ Kurtosis IQR	**1.00**	0.00	0.00	0.00
$STFT_{pow}$ Skewness IQR	**0.95**	0.08	–0.30	0.02
Bark Spread Median	**0.82**	0.57	–0.02	–0.11
$STFT_{mag}$ Decrease Median	**0.78**	0.48	–0.40	0.02
Bark Crest Median	**0.76**	0.61	0.23	–0.03
Harmonic Kurtosis IQR	**0.76**	–0.13	–0.19	–0.61
$STFT_{pow}$ Frame erg IQR	–0.00	**1.00**	0.00	–0.00
Harmonic Frame erg IQR	0.46	**0.82**	0.20	–0.28
Effective Duration	–0.44	**0.80**	0.36	–0.21
Harmonic Decrease Median	0.10	**0.76**	0.16	0.62
$STFT_{mag}$ Flatness IQR	–0.00	–0.00	**1.00**	0.00
$STFT_{pow}$ Crest IQR	0.19	–0.21	**0.94**	0.19
$STFT_{mag}$ Crest IQR	–0.00	0.00	–0.00	**1.00**
$STFT_{mag}$ Centroid IQR	0.67	0.31	–0.47	0.48
$STFT_{pow}$ Kurtosis Median	0.69	0.42	0.13	0.58
$STFT_{pow}$ Skewness Median	–0.18	0.71	–0.64	–0.22
Bark Centroid Median	0.66	0.72	0.09	0.18

Note: Bold type signifies absolute component loading >0.75. Features with loadings at this level are used to label components, as in [26]. Frame erg, framewise energy; IQR, inter-quartile range; $STFT_{pow}$, power spectrogram from short time Fourier transform; $STFT_{mag}$, magnitude spectrogram from short time Fourier transform.

nent loadings are shown in Table 4. Features with loadings above a threshold (set at 0.75) are used to label components.

The first component showed above-threshold loadings for the medians of spectral decrease [52], Bark spectral spread, and crest factor. It also showed above-threshold loadings for IQRs of the skewness and kurtosis of the STFT power spectrum and harmonic magnitudes. This somewhat contradictory combination of spectral features implies this component describes a continuum between specific spectrotemporal profiles. The second component shows above-threshold loadings for median harmonic decrease and the IQRs of frame energies in both the STFT power and harmonic magnitude spectra. It also showed a positive loading for effective duration. These loadings suggest this component describes a sound with a longer sustain and high temporal energy variation.

The third component shows above-threshold loadings for the IQRs of STFT magnitude flatness and STFT power crest factor. These loadings imply that a sound with a high score on this component would contain spectrotemporal modulations that vary between a flat spectral distribution (typically indicative of a noisy or inharmonic sound) and a spectrum with a distinct crest. This may suggest that sounds with a high loading on this component may be more likely to make use of the amplitude envelopes of the synthesizer's modulating operators. The final component shows an above-threshold loading for the IQR of STFT magnitude crest factor. This suggests that sounds with a high score on this component may, again, employ the amplitude envelopes of the modulating operators in a way that moves between a pronounced spectral peak and a more even energy distribution.

Table 5 shows Spearman rank correlation coefficients between the five semantic factors and the four acoustic components. To accommodate the comparative nature of the semantic ratings, analysis was performed using the difference between the created sound and its reference along each acoustic component. In interpreting these coefficients and their significance, it is important to take into account the large number of sounds in this analysis ($n = 270$), inherent noise in the dataset caused by the single rating provided for each sound, and subjectivity of assigning a value to the applicability of a semantic descriptor. In particular, while many correlations were significant at the $p < 0.001$ level, the strengths of their relationships were moderate.

The first factor (*sharpness*) showed significant negative correlations with components PC1, PC2, and PC4 and a significant positive correlation with component PC3. Factors F3–F5 all share a pattern of highly significant correlations with components PC1 and PC3, with factor F3 inverted compared to the other two. The second factor (associated with *mass*) did not show significant correlations with any of the principal components of acoustic variation. Similarly there was no influence of stimulus F0 on any of the semantic factors.

2.3 Synthesizer Parameters

The perceptual imprints of timbre on the sound design process were inspected next. In order to identify whether semantic prompts and the direction of comparison exerted an effect on the adjustments made to synthesizer controls, linear regression models were computed for every $\Delta(p_c - p_r)$ and F0 with comparative prompt as a categorical variable with six levels, i.e., three adjectives in two directions of comparison. Estimated regression slopes (β coefficients) served as indicators of effect size (see Fig. 5).

Similar patterns of linear effects on changes to the modulator tuning and volume parameters for *brighter*, *less bright*,

Table 5. Spearman rank correlation coefficients between semantic factors and acoustic feature principal components, as well as fundamental frequency.

	PC1 Spectrotemporal (Distribution) & Spectral Shape	PC2 Temporal Energy Variation & Spectral Slope	PC3 Spectrotemporal (Flatness)	PC4 Spectrotemporal (Crest Factor)	F0
Factor 1 (Sharpness)	–0.58***	–0.37***	0.49***	–0.25***	–0.01
Factor 2 (Mass)	0.09	–0.02	0.09	0.03	0.08
Factor 3 (Clarity)	0.29***	0.17**	–0.44***	0.04	–0.03
Factor 4 (Percussiveness)	–0.24***	–0.03	0.31***	–0.14*	–0.02
Factor 5 (Rawness)	–0.22***	–0.10	0.34***	–0.10	–0.05

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

and *less rough* prompts were observed, with the polarity of the effects inverted for the "less" prompts. These effects were also present for *rougher*, although they are less pronounced. Given the properties of FM synthesis, these similarities are intuitive: these parameters directly dictate the intensity, energy distribution, and partial distribution of the modulated signal.

The *more thick* prompt showed consistent effects on the amplitude envelope controls of both the carrier and modulating operators. This suggests that thickness is modulated by manipulating both the sustain of overall amplitude and sustain of sideband energy. However the width of the 95% confidence intervals of these effects implies a large degree of variance in how these controls were actually used in re-

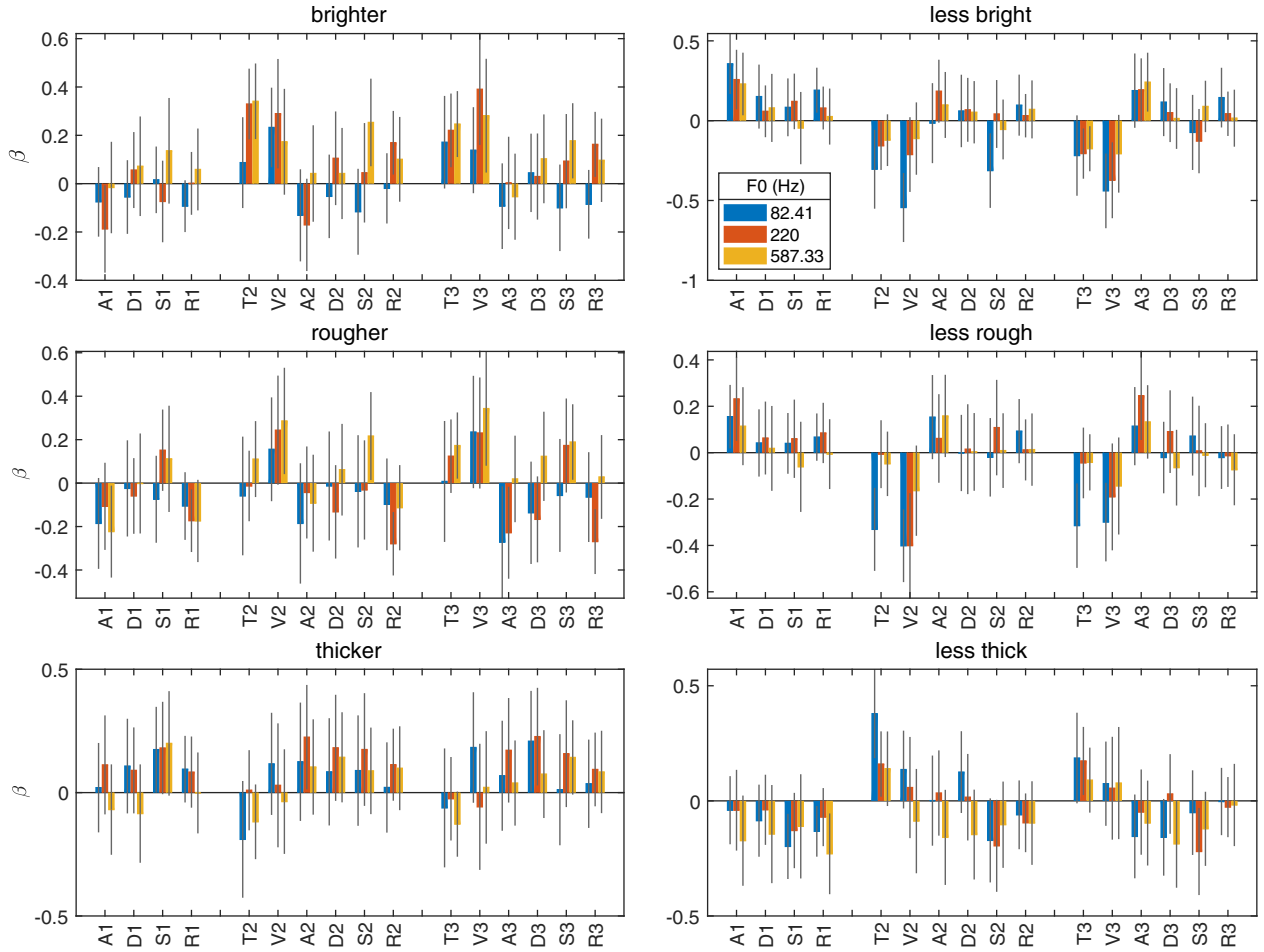

Fig. 5. Linear effects (β) of comparative semantic prompt derived from linear regression for every frequency-modulation synthesizer control change and fundamental frequency. Error bars correspond to 95% confidence intervals. A = attack; D = decay; R = release; S = sustain; T = tuning; V = volume; 1 = carrier; 2/3 = modulators.

sponse to prompts. In the case of modulator controls, this may be explained by their equivalence in the architecture of the synthesizer—that is, swapping the control values of operators 2 and 3 results in an identical sound being produced. Achieving a change in accordance with a given prompt may therefore not require the manipulation of all controls capable of achieving changes along that semantic dimension, thus weakening the statistical relationship between each such control and its corresponding prompt.

In general, prompt effects on tuning and volume controls were observed to be consistently stronger than on ADSR envelope controls. This may be partially because of the interdependence of synthesis parameters—the strength and nature of the effect of the ADSR parameters of a modulating operator are dictated by the values of the corresponding tuning and volume controls. For example, if the volume control of an operator is very low, the strength of the effect of the envelope sustain control may be almost imperceptible. However the weak ADSR effects are probably mostly because of the lack of a prompt that explicitly describes temporal characteristics of a signal. Because an explicit percussive-plucky factor emerged in the analysis of post-hoc semantic ratings, such a prompt would be a useful addition to future applications of this prompted synthesis paradigm.

To examine the relationship between adjustments to synthesizer controls, semantic factors, and the principal axes of acoustical variation, Spearman's rank correlation coefficient was computed between synthesizer control changes $\Delta(p_c - p_r)$, semantic factor scores, and differences between created/reference sounds along acoustic principal components. These values are displayed in Fig. 6. Correlations were generally strongest across all factors for the tuning and volume controls of the modulating operators, suggesting these exerted a larger influence over both semantic ratings and the resulting acoustic properties of synthesized sounds. Modulator volume, however, appeared to exhibit almost no relationship with factor F2 (*mass*), while correlating significantly with all other semantic factors and acoustic principal components. This may imply that the concept of semantic mass is less significantly influenced by the sideband energy in the signal.

Comparatively correlations with ADSR envelope controls were generally weaker, although the carrier operator's attack control showed moderately strong inverse relationships with factors F1 (*sharpness*), F4 (*percussive*), and F5 (*rawness*). Modulator attack controls also showed moderate negative correlations with F4 suggesting, as might be expected from musical intuition, that greater percussiveness is characterized by both a shorter attack portion in the amplitude envelope with a short transient with a wider spectral distribution. Again the weaker relationships seen in other envelope controls may have arisen because of the lack of a specifically temporal prompt descriptor.

3 DISCUSSION

The semantic correspondences of a wide variety of sounds, produced through FM synthesis, were explored using a novel experimental paradigm based on a prompted synthesis task. Experienced sound designers both created sounds in response to prompts and provided semantic ratings on the sounds they produced. These responses were studied by constructing a semantic timbre space using exploratory factor analysis, and a correlation analysis was performed with the principal components of a set of acoustic features. Finally the influence of semantic prompts on the sound design process was examined by fitting linear models to synthesizer parameter changes.

The five-factor semantic space for FM sounds identified by the analysis in the previous section showed strong loadings for timbral descriptions associated with the LTM dimensions observed previously for acoustic and electroacoustic instrument tones [26, 7] but also exhibited a distinct structure in response to the specificities of FM signals. The recurrence of LTM-like factors in this and previous studies indicates that these concepts may generalize well across timbral domains, while the occurrence of more highly specified factors suggests that these concepts alone do not form a complete timbre semantic model. In interpreting these results, it is crucial to be mindful that these observations cannot be assumed to generalize beyond the timbral domain of the experimental FM synthesizer. Continued inquiry into the full diversity of electronic sound is needed to understand the extent to which the findings are because of specificities of FM synthesis.

3.1 Implications for the Perception and Semantic Processing of Timbre

The first factor, which was labeled *sharpness*, showed strong loadings for both luminance-related and texture-related words, although less so for *rough* and *smooth*, suggesting it may represent an amalgam of attributes relating to these two semantic dimensions. It has been suggested that a *sharp* timbre is one that is both *bright* and *rough* [53]. The acoustic principal component correlates of F1 were the strongest seen across all five factors, suggesting it may be more closely related than other factors to the main aspects of acoustic variation in the created sounds. This was also the case for the musical timbres investigated in [26], where, albeit separately, the two luminance and texture factors shared their most significant acoustic correlations.

In the context of FM synthesis, where the introduction of brightness (in the form of high-frequency energy) is closely linked to the introduction of inharmonicity through phase modulation, an entanglement of luminance and texture may follow naturally. Thus the closer alignment of these two semantic concepts in this study could be a direct result of the chosen method of synthesis. The similarities between the effects of *bright* and *rough* prompts on modulator volume and tuning synthesizer controls (Fig. 5) might further support this interpretation. That is to say, the same controls were used when participants were asked to modulate the perceived brightness as when they were asked to decrease the perceived roughness. However prompts to increase roughness did not result in quite so strong an effect, suggesting there may exist a degree of independence be-

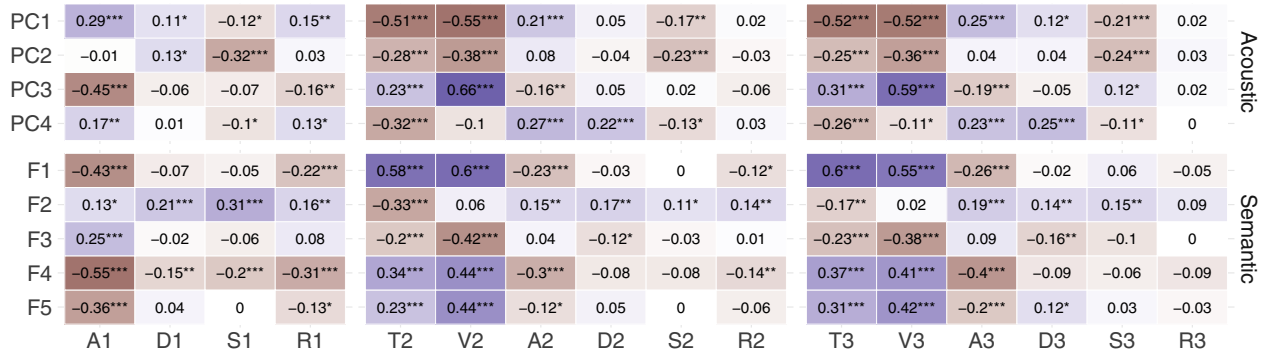

Fig. 6. Spearman's ρ computed between changes to synthesizer controls, semantic factors, and differences along acoustic principal components. A = attack; D = decay; F1–F5 = semantic factors; PC1–PC4 = acoustic principal components; R = release; S = sustain; T = tuning; V = volume; 1 = carrier; 2 & 3 = modulators.

tween brightness and roughness that could not be entirely captured by the factor model.

Acoustic correlations for the second semantic factor (*mass*) were less clear. On the one hand, this might be the result of the acoustical analysis lacking an audio descriptor or a set of descriptors that adequately capture the concept of sound mass. Alternatively it is plausible that a number of possible combinations of characteristics independently associate with auditory mass, and the scale and structure of the dataset has obscured any such individual correlations. Indeed the two most highly correlated acoustic principal components (PC1 and PC3) described changes in the shape and flatness of the spectral distribution over time, which might suggest that the semantic dimensions of this set of FM synthesizer sounds are best characterized by modulation of these spectrotemporal characteristics. Recent work shows that spectrotemporal modulation representations could explain a higher amount of the variance in semantic ratings of sound mass than classical audio descriptors of the type used here [54].

The third (with strong loadings for *clean* and *clear*) and fifth (with a strong loading for *raw*) factors described more nuanced aspects of timbral variation, specific to FM-synthesized sounds. FM synthesis provides fine-grained control over the distribution of partials, with the energy distribution over sidebands governed by Bessel functions of the modulation index [37]. It is plausible that certain aspects of variation between FM-synthesized sounds are pronounced enough to be differentiated by similarly fine-grained semantic dimensions and may otherwise be less separable in other contexts. For instance, in the LTM study, English listeners perceived *messy* acoustic and electroacoustic instrument tones to also be *rough* and, to a lesser extent, *thick*, while scales like *clear* and *dirty* were dropped from the final factor analysis because of high correlation with other scales [26].

On the other hand, the emergence of a *plucky/percussive* dimension (factor F4) in the present study might be interpreted from a methodological angle. Interacting with the synthesizer's ADSR envelopes may have encouraged participants, who also had significant prior sound design experience, to be particularly sensitive to the temporal shape of the sounds they actively created, where they might not be in a conventional passive listening design. Indeed factors proposed across several such investigations of timbre semantics, including the LTM study, appear generally unable to capture the salient perceptual dimension of timbre responsible for discriminating between sustained and impulsive sounds [7, 27].

While this factor shows weak to moderate correlations with some acoustic components, no relationship was observed with the only component (PC2) associated with a descriptor related to temporal energy (effective duration). It is possible that, in the context of FM-synthesized sounds, the attributes insinuated by the terms *percussive* and *plucky* are not well characterized by purely temporal descriptors. These terms may, for example, be more suggestive of particular profiles of spectrotemporal evolution. They are also distinct from other semantic descriptors in both this analysis and previous work [18, 26] because, instead of being metaphors for timbral characteristics, they may be directly suggestive of source-cause categorical cues such as striking and plucking. Timbrally, these are typically associated with an instantaneous attack transient, after which the signal energy decays. It stands to reason then that the inclusion of *percussive* and *plucky* scales might have been sufficient to elicit discrimination of such timbral characteristics, despite this not being a principal component of acoustic variation.

3.2 Relationship Between Semantic Factors and Synthesis Parameters

Significant correlations between the observed semantic factors and adjustments made by participants to synthesis parameters were observed (Fig. 6). In order to interpret these correlations, it is helpful to understand how the parameters of an FM synthesizer influence the resulting signal at a high level. Thus the authors propose conceptually dividing the parameters of the synthesizer into the following four groups, based on their effects:

1. Amplitude temporal evolution: carrier attack (A1), decay (D1), sustain (S1), and release (R1).
2. Spacing between sideband frequencies: modulator tuning (T2 and T3).

3. Sideband energy distribution: modulator volume (V2 and V3).
4. Sideband energy temporal evolution: modulator attack (A2 and A3), decay (D2 and D3), sustain (S2 and S3), and release (R2 and R3).

With these groupings in mind, analyzing the pattern of correlations seen for each factor becomes a simpler task. Increasing "sharpness" (F1) appears, for example, to be associated with (1) faster amplitude envelopes (↓A1 and ↓R1), (2) wider spacing between sidebands (↑T2 and ↑T3), (3) more energy distributed to sidebands (↑V2 and ↑V3), and (4) a shorter sideband energy envelope (↓A2 and ↓A3). Conversely increasing "mass" (F2) suggests parameter changes that cause (1) slower amplitude envelopes with more sustain (↑D1, ↑S1, and ↑R1), (2) narrower spacing between sidebands (↓T2 and ↓T3), (3) no change to sideband energy distribution, and (4) slower sideband energy envelopes with more sustain (↑A2, ↑D2, ↑R2, ↑A3, ↑D3, and ↑S3).

Through this lens, the semantic factor/synthesis-parameter relationships are somewhat intuitive. Percussiveness (F4) is mostly associated, for example, with shorter envelopes and more energy in sidebands, which is consistent with previous definitions of "percussive" semantic dimensions [27]. However many of the semantic factor/synthesis-parameter correlations are statistically significant but exhibit only a small correlation, which is congruent with the high variance also seen in parameter changes per prompt. This suggests that, as with the prompt-parameter relationships, the distribution of semantic factor/synthesis-parameter relationships is highly varied and exhibits nuances likely resulting from the specifics of FM synthesis discussed in the following section.

3.3 Influence of Task Constraints and Pitch Register

More generally the hands-on synthesis component of the present experiment may have resulted in heightened sensitivity to certain timbral cues, such as those captured by factors 3–5. These, although commonly shared across many types of sounds, may be more difficult to perceptually disentangle in complex natural versus simple synthetic sounds (see, for example, [55]). As such, the latter may have invited for subtler semantic associations. Reusing previously created sounds as reference stimuli for each trial may have also contributed to the prominence of timbral subtleties in the factor space. Given the greater diversity of stimuli included in the analysis, it is reasonable to assume that a wider diversity of sonic characteristics were represented. However, because each stimulus pair was rated only once, it is not possible to quantify inter-rater agreement on the presence or distribution of these characteristics. It would therefore be beneficial, in future work, to collect semantic ratings from multiple participants on a shared set of stimuli similar to those used in this study.

Another methodological choice that might have driven the finer-grained factor solution is the use of pairwise comparative ratings, which are generally considered to not limit the dimensionality that can be recovered [55]. Because participants rated semantic scales based on the dissimilarity between a reference sound and the one they created, one stimulus pair at a time, differentiating timbral subtleties that may be obscured in an absolute rating paradigm might have been enabled (although see [56]). Further work collecting absolute semantic ratings on the same stimuli would be necessary to confirm this.

The comparative nature of the semantic ratings might also explain the lack of any significant relationship between stimulus F0 and the five semantic factors in the present data. At first this finding would appear at odds with previous reports both when F0/pitch is examined directly [41, 42] and when it is considered as an additional variable [26]. In the LTM space, for instance, F0 was found strongly correlated with the mass dimension, with lower-pitched sounds rated as thicker and more dense (cf. [57]). It is possible that the use of comparative versus absolute rating scales effectively controlled for any F0 effects. Another plausible explanation is that the specific characteristics of FM synthesis may have perceptually obscured the true F0 of some sounds. That is, the introduction of sidebands both above and below the oscillating frequency of the carrier operator might have falsely implied a lower or higher pitch [58].

The architecture of the FM synthesizer used by participants may have limited the power of the linear models presented in Sec. 2.3 to accurately predict the influence of semantic prompts on parameter changes. In particular the symmetry of the modulation routing means that swapping the parameter values of operators 2 and 3 would result in an identical sound being produced. This is reflected in the similarity of the linear effects (Fig. 5) between the parameters of both modulators and may have weakened the statistical relationships between modulator parameters and semantic descriptors.

There also exist degenerate regions in the synthesis-parameter space, such as when the amplitude of a modulating operator is zero. In these cases, none of the parameters of the modulator in question contribute to the resulting audio signal, although still influencing the statistical analysis. Future applications of this paradigm, therefore, would benefit from either an asymmetric synthesis architecture or an analysis that accounts for parameter redundancies and degeneracies. Further experimentation with a linear synthesis method, such as additive synthesis, would also help understand to what extent these results derive from the nonlinearity and complexity of FM synthesis.

Furthermore a given semantic prompt may not map uniquely to a single point in the synthesizer's parameter space as per the instructed task. This is because of both the previously discussed symmetry of the synthesizer and the fact that the synthesizer's parameters may not map directly onto the semantic dimensions under test. For example, it is plausible that the neighborhood surrounding a "bright" sound in the parameter space also consists largely of "bright" sounds. It is also conceivable that there may exist several disjoint neighborhoods in parameter space that satisfy a "bright" timbre. As such, the collected data may

represent an incomplete picture of a listener's belief about the distributions of semantic descriptors across the synthesizer's parameter space because they provide only point estimates. Further research aiming to map these distributions across the ranges of parameters would therefore be valuable.

3.4 Influence of Word Affect on Timbre-Semantic Associations

In the prompted synthesis study of Wallmark et al. [40], affective connotations of the adjective prompts (based on validated affect norms [59]) were found to exert an influence over the acoustic properties of the created sound. Words with positive or negative valence were observed to result in higher scores on an acoustic component associated with spectral centroid and noisiness. Words with neutral valence, conversely, were associated with lower scores on this component.

Largely similar trends were observed for the FM sounds created in response to the three prompts used in the present study, which respectively have positive valence (*bright*), neutral valence (*thick*), and negative valence (*rough*) [59]. Specifically the patterns of linear effects in Fig. 5 indicate that the largest effects for *brighter*, *less bright*, and *less rough* were on the tuning and volume controls of the two modulators, albeit with some inconsistency between pitch registers; *thicker* and *less thick* showed overall weaker linear effects for the same controls. These controls were strongly associated with both spectral centroid (PC1) and noisiness (PC3; Table 4 and Fig. 6). While a systematic examination of the acoustical impact of word affect remains beyond the scope of this paper, the present data provide additional preliminary evidence of affective mediation in timbre semantics.

3.5 Toward Perceptually Informed Sound Design and Synthesis

As observed in the present study and in previous work [8], the controls of existing synthesizers generally do not provide a clear mapping onto timbral concepts. Broadly speaking, they instead map onto specifics of the underlying synthesis method requiring musicians and sound designers to acquire some level of signal processing knowledge in order to make principled decisions. Even with this knowledge, achieving conceptually simple alterations often requires manipulation of multiple parameters, often in a counter-intuitive manner governed by their subtle interdependence. This issue is further compounded by the growing complexity of commercial hardware and software synthesizers.

Wessel [9] first suggested the use of a timbre dissimilarity space, constructed using multidimensional scaling, as a control space for a synthesizer. The proposed approach used an additive synthesis engine whose envelope parameters were mapped linearly to the dimensions of the timbre space. Such a simple mapping was likely facilitated by the linearity of additive synthesis, where the signal is constructed as a time-varying weighted sum of a set of basis functions. FM synthesis, conversely, constructs a signal from synthesis parameters nonlinearly, and many controls are thus arguably "perceptually nonlinear." For example, monotonically increasing a modulator's frequency parameter over time would result in a signal that oscillates between harmonic structure and total inharmonicity. Thus simple timbre space mappings to FM parameters can be more challenging to derive [60, 61]. Furthermore mapping synthesis parameters to a semantic timbre space introduces yet another layer of complexity because, while timbre-semantic dimensions are assumed to relate to an underlying perceptual representation, the nature of this relationship is not clear for all dimensions [18, 62].

As research in neural audio synthesis [63] extends the capabilities of synthesizers beyond the limitations of familiar techniques, a further set of challenges related to synthesis control warrants consideration. It is now already feasible to create convincing digital recreations of the sounds of physical musical instruments without the need for sample playback or physical modeling [64], transfer the timbre of one instrument to another [65], perform perceptually smooth "morphs" between timbres [10], and more. Recent work [66] has enabled many of these techniques to be achieved comfortably in real time on consumer central processing units, allowing the capabilities of neural audio synthesis to be integrated into tools for musicians and sound designers. Yet affording useful timbral control over these tools remains an unsolved problem. Their range of potential outputs is huge, yet their internal representations of timbral characteristics are typically learned directly from training data and are frequently uninterpretable by humans.

Yet, without a complete understanding of how synthetic sounds are perceived, which characteristics are most perceptually salient, how this perception maps onto comprehensible descriptions, and how these descriptions guide the sounds design process, such work is unlikely to produce controls of practical utility to those hoping to exploit the vast sonic potential of these new synthesizers in their creative work. Previous work has focused on addressing this problem in the context of audio engineering and music production by studying the relationships between semantic descriptors of timbre and the application of audio effects including equalization, compression, reverb [34], distortion [35], and bit-depth reduction [67]. Progress on this problem for audio synthesis will require interdisciplinary collaboration across the fields of psychoacoustics, deep learning, and human-computer interaction. To this end, this work is accompanied by a fully annotated dataset of sounds produced in the study, with complete semantic ratings and factor loadings. This is intended as a first step toward sharing insights across these fields in a manner that will facilitate progress on this problem.

4 CONCLUSION

This study investigated the semantic associations of disembodied electronic timbres—specifically those produced by a three-operator FM synthesizer. A novel experimental paradigm was applied in which participants directly syn-

thesized sounds in response to semantic prompts linked to the dimensions of the LTM model of timbre semantics. An exploratory factor analysis of comparative semantic ratings collected between pairs of synthesized sounds recovered a five-factor semantic space. To identify the acoustic underpinnings of the resulting factors, a correlation analysis was performed with the principal components of a comprehensive set of acoustic features. Linear regression models were also fit to examine the effects of semantic prompts on the use of synthesizer controls.

Semantic factors corresponding to *luminance*, *texture*, and *mass* were present in this model, but *luminance* and *texture* were combined. Acoustic correlates of *luminance* and *texture* similar to those observed in previous work [26] were found, but no acoustic correlates could be directly identified for *mass*. Three additional factors were observed with no obvious parallel in the LTM model. These showed strong loadings for *clear/clean*, *percussive/plucky*, and *raw*, respectively. No influence of fundamental frequency on the ratings of semantic descriptors was observed, likely because of their comparative nature. All three comparative LTM prompts exerted significant influence on the manipulation of synthesizer controls. The prompts *brighter*, *less bright*, and *less rough* in particular were very significantly associated with changes to parameters directly controlling the FM modulation index. Future work aiming to ascertain the nature of this model's three novel dimensions would be valuable. The application of classical timbre dissimilarity and semantics paradigms to sounds generated in this study would also facilitate interpretation of these results in the broader context of timbre research.

5 ACKNOWLEDGMENT

This work was supported by UK Research and Innovation [grant number EP/S022694/1]. C.S. thanks Asterios Zacharakis for fruitful discussions and methodological recommendations.

6 REFERENCES

[1] 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 (Eds.), *Timbre: Acoustics, Perception, and Cognition*, Springer Handbook of Auditory Research, vol. 69, pp. 1–19 (Springer, Cham, Switzerland, 2019). https://doi.org/10.1007/978-3-030-14832-4_1.

[2] C. Fales, "Hearing Timbre: Perceptual Learning Among Early Bay Area Ravers," in R. Fink, M. Latour, and Z. Wallmark (Eds.), *The Relentless Pursuit of Tone: Timbre in Popular Music*, pp. 21–42 (Oxford University Press, New York, NY, 2018). https://doi.org/10.1093/oso/9780199985227.003.0002.

[3] S.-A. Lembke, "Hearing Triangles: Perceptual Clarity, Opacity, and Symmetry of Spectrotemporal Sound Shapes," *J. Acoust. Soc. Am.*, vol. 144, no. 2, pp. 608–619 (2018 Aug.). https://doi.org/10.1121/1.5048130.

[4] C. Vahidi, G. Fazekas, C. Saitis, and A. Palladini, "Timbre Space Representation of a Substractive Synthesiser," in *Proceedings of the 2nd International Conference on Timbre*, pp. 30–33 (Thessaloniki, Greece) (2020 Sep.).

[5] T. Grill, A. Flexer, and S. Cunningham, "Identification of Perceptual Qualities in Textural Sounds Using the Repertory Grid Method," in *Proceedings of the 6th Audio Mostly Conference: A Conference on Interaction With Sound*, pp. 67–74 (Coimbra, Portugal) (2011 Sep.). https://doi.org/10.1145/2095667.2095677.

[6] M. Carron, T. Rotureau, F. Dubois, N. Misdariis, and P. Susini, "Speaking About Sounds: A Tool for Communication on Sound Features," *J. Des. Res.*, vol. 15, no. 2, pp. 85–109 (2017 Sep.). https://doi.org/10.1504/JDR.2017.086749.

[7] C. Saitis and S. Weinzierl, "The Semantics of Timbre," in K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay (Eds.), *Timbre: Acoustics, Perception, and Cognition*, Springer Handbook of Auditory Research, vol. 69, pp. 119–149 (Springer, Cham, Switzerland, 2019). https://doi.org/10.1007/978-3-030-14832-4_5.

[8] A. Seago, S. Holland, and P. Mulholland, "A Critical Analysis of Synthesizer User Interfaces for Timbre," in *Proceedings of the 18th British HCI Group Annual Conference*, vol. 2, pp. 105–108 (Leeds, UK) (2004 Sep.).

[9] D. L. Wessel, "Timbre Space as a Musical Control Structure," *Comput. Music J.*, vol. 3, no. 2, pp. 45–52 (1979 Jun.). https://doi.org/10.2307/3680283.

[10] P. Esling, A. Chemlag–Romeu-Santos, and A. Bitton, "Bridging Audio Analysis, Perception and Synthesis With Perceptually-Regularized Variational Timbre Spaces," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pp. 175–181 (Paris, France) (2018 Sep.).

[11] S. Le Groux and P. F. Verschure, "Perceptsynth: Mapping Perceptual Musical Features to Sound Synthesis Parameters," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 125–128 (Las Vegas, NV) (2008 Mar.). https://doi.org/10.1109/ICASSP.2008.4517562.

[12] S. Conan, E. Thoret, M. Aramaki, et al., "An Intuitive Synthesizer of Continuous-Interaction Sounds: Rubbing, Scratching, and Rolling," *Comput. Music J.*, vol. 38, no. 4, pp. 24–37 (2014 Dec.). https://doi.org/10.1162/COMJ_a_00266.

[13] M. Aramaki, M. Besson, R. Kronland-Martinet, and S. Ystad, "Controlling the Perceived Material in an Impact Sound Synthesizer," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 2, pp. 301–314 (2011 Feb.). https://doi.org/10.1109/TASL.2010.2047755.

[14] H. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, A. J. Ellis (Trans.) (Dover Publications, Mineola, NY, 1954), 2nd ed.

[15] R. Plomp, "Timbre as a Multidimensional Attribute of Complex Tones," in R. Plomp and G. F. Smoorenburg (Eds.), *Frequency Analysis and Periodicity Detection in Hearing*, pp. 397–410 (Sijthoff, Leiden, Netherlands, 1970).

[16] J. M. Grey, "Multidimensional Perceptual Scaling of Musical Timbres," *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1270–1277 (1977 May). https://doi.org/10.1121/1.381428.

[17] 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," *Psychol. Res.*, vol. 58, no. 3, pp. 177–192 (1995 Dec.). https://doi.org/10.1007/BF00419633.

[18] T. M. Elliott, L. S. Hamilton, and F. E. Theunissen, "Acoustic Structure of the Five Perceptual Dimensions of Timbre in Orchestral Instrument Tones," *J. Acoust. Soc. Am.*, vol. 133, no. 1, pp. 389–404 (2013 Jan.). https://doi.org/10.1121/1.4770244.

[19] E. Thoret, B. Caramiaux, P. Depalle, and S. McAdams, "Learning Metrics on Spectrotemporal Modulations Reveals the Perception of Musical Instrument Timbre," *Nat. Hum. Behav.*, vol. 5, pp. 369–377 (2020 Nov.). https://doi.org/10.1038/s41562-020-00987-5.

[20] C. E. Osgood, "The Nature and Measurement of Meaning," *Psychol. Bull.*, vol. 49, no. 3, pp. 197–237 (1952 May). https://doi.org/10.1037/h0055737.

[21] R. A. Kendall and E. C. Carterette, "Verbal Attributes of Simultaneous Wind Instrument Timbres: I. von Bismarck's Adjectives," *Music Percept.*, vol. 10, no. 4, pp. 445–467 (1993 Jul.). https://doi.org/10.2307/40285583.

[22] L. N. Solomon, "Semantic Approach to the Perception of Complex Sounds," *J. Acoust. Soc. Am.*, vol. 30, no. 5, pp. 421–425 (1958 May). https://doi.org/10.1121/1.1909632.

[23] G. von Bismarck, "Timbre of Steady Sounds: A Factorial Investigation of Its Verbal Attributes," *Acta Aust. united Acust.*, vol. 30, no. 3, pp. 146–159 (1974 Mar.).

[24] R. L. Pratt and P. E. Doak, "A Subjective Rating Scale for Timbre," *J. Sound Vib.*, vol. 45, no. 3, pp. 317–328 (1976 Apr.). https://doi.org/10.1016/0022-460X(76)90391-6.

[25] A. C. Disley, D. M. Howard, and A. D. Hunt, "Timbral Description of Musical Instruments," in *Proceedings of the 9th International Conference of Music Perception and Cognition*, pp. 61–68 (Bologna, Italy) (2006 Aug.).

[26] A. Zacharakis, K. Pastiadis, and J. D. Reiss, "An Interlanguage Study of Musical Timbre Semantic Dimensions and Their Acoustic Correlates," *Music Percept.*, vol. 31, no. 4, pp. 339–358 (2014 Apr.). https://doi.org/10.1525/mp.2014.31.4.339.

[27] A. Zacharakis and K. Pastiadis, "Revisiting the Luminance-Texture-Mass Model for Musical Timbre Semantics: A Confirmatory Approach and Perspectives of Extension," *J. Audio Eng. Soc.*, vol. 64, no. 9, pp. 636–645 (2016 Sep.). https://doi.org/10.17743/jaes.2016.0032.

[28] L. Reymore and D. Huron, "Using Auditory Imagery Tasks to Map the Cognitive Linguistic Dimensions of Musical Instrument Timbre Qualia," *Psychomusic.: Music Mind Brain*, vol. 30, no. 3, pp. 124–144 (2020 Jun.). https://doi.org/10.1037/pmu0000263.

[29] A. Zacharakis and J. Reiss, "An Additive Synthesis Technique for Independent Modification of the Auditory Perceptions of Brightness and Warmth," presented at the *130th Convention of the Audio Engineering Society* (2011 May), paper 8420.

[30] C. Saitis, K. Siedenburg, P. M. Schuladen, and C. Reuter, "The Role of Attack Transients in Timbral Brightness Perception," in *Proceedings of the 23rd International Congress on Acoustics*, paper 5506 (Aachen, Germany) (2019 Sep.).

[31] R. Ethington and B. Punch, "SeaWave: A System for Musical Timbre Description," *Comput. Music J.*, vol. 18, no. 1, pp. 30–39 (1994 Spring). https://doi.org/10.2307/3680520.

[32] A. Gounaropoulos and C. Johnson, "Synthesising Timbres and Timbre-Changes from Adjectives/Adverbs*Applications of Evolutionary Computing*, Lecture Notes in Computer Science, vol. 3907, pp. 664–675 (Springer, Berlin, Germany, 2006). https://doi.org/10.1007/11732242_63.

[33] M. Cartwright and B. Pardo, "Social-EQ: Crowdsourcing an Equalization Descriptor Map," in *Proceedings of the 14th International Society for Music Information Retrieval Conference*, pp. 395–400 (Curitiba, Brazil) (2013 Nov.).

[34] R. Stables, S. Enderby, B. De Man, G. Fazekas, and J.D. Reiss, "SAFE: A System for Extraction and Retrieval of Semantic Audio Descriptors," presented at the *15th International Society for Music Information Retrieval Conference* (Taipei, Taiwan) (2014 Oct.).

[35] R. Stables, S. Enderby, et al., "Semantic Description of Timbral Transformations in Music Production," in *Proceedings of the 24th ACM International Conference on Multimedia*, pp. 337–341 (Amsterdam, Netherlands) (2016 Oct.). https://doi.org/10.1145/2964284.2967238.

[36] P. Esling, N. Masuda, A. Bardet, R. Despres, and A. Chemla-Romeu-Santos, "Flow Synthesizer: Universal Audio Synthesizer Control With Normalizing Flows," *Appl. Sci.*, vol. 10, no. 1, paper 302 (2020 Jan.). https://doi.org/10.3390/app10010302.

[37] J. M. Chowning, "The Synthesis of Complex Audio Spectra by Means of Frequency Modulation," *J. Audio Eng. Soc.*, vol. 21, no. 7, pp. 526–534 (1973 Sep.).

[38] D. Wessel, D. Bristow, and Z. Settel, "Control of Phrasing and Articulation in Synthesis," in *Proceedings of the International Computer Music Conference*, pp. 108–116 (Champaign, IL) (1987 Aug.).

[39] R. D. Ashley, "A Knowledge-Based Approach to Assistance in Timbral Design," in *Proceedings of the International Computer Music Conference*, pp. 11–16 (The Hague, Netherlands) (1986 Oct.).

[40] Z. Wallmark, R. J. Frank, and L. Nghiem, "Creating Novel Tones From Adjectives: An Exploratory Study Using FM Synthesis," *Psychomusic.: Music Mind Brain*, vol. 29, no. 4, pp. 188–199 (2019 Jul.). https://doi.org/10.1037/pmu0000240.

[41] K. M. Steele and A. K. Williams, "Is the Bandwidth for Timbre Invariance Only One Octave?" *Music Percept.*, vol. 23, no. 3, pp. 215–220 (2006 Feb.). https://doi.org/10.1525/mp.2006.23.3.215.

[42] J. Marozeau and A. de Cheveigné, "The Effect of Fundamental Frequency on the Brightness Dimension of

Timbre," *J. Acoust. Soc. Am.*, vol. 121, no. 1, pp. 383–387 (2007 Jan.). https://doi.org/10.1121/1.2384910.

[43] 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, paper e89642 (2014 Feb.). https://doi.org/10.1371/journal.pone.0089642.

[44] 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*, pp. 129–132 (Thessaloniki, Greece) (2020 Sep.).

[45] F. Henninger, Y. Shevchenko, U. Mertens, P. J. Kieslich, and B. E. Hilbig, "lab.js: A Free, Open, Online Experiment Builder," *Zenodo* (2020 Apr.). https://doi.org/10.5281/zenodo.3767907.

[46] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan, "Evaluating the Use of Exploratory Factor Analysis in Psychological Research," *Psychol. Methods*, vol. 4, no. 3, pp. 272–299 (1999 Sep.). https://doi.org/10.1037/1082-989X.4.3.272.

[47] H. F. Kaiser, "The Application of Electronic Computers to Factor Analysis," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 141–151 (1960 Apr.). https://doi.org/10.1177/001316446002000116.

[48] R. B. Cattell, "The Scree Test for the Number of Factors," *Multivar. Behav. Res.*, vol. 1, no. 2, pp. 245–276 (1966 Apr.). https://doi.org/10.1207/s15327906mbr0102_10.

[49] J. L. Horn, "A Rationale and Test for the Number of Factors in Factor Analysis," *Psychometrika*, vol. 30, no. 2, pp. 179–185 (1965 Jun.). https://doi.org/10.1007/BF02289447.

[50] W. R. Zwick and W. F. Velicer, "Comparison of Five Rules for Determining the Number of Components to Retain," *Psychol. Bull.*, vol. 99, no. 3, pp. 432–442 (1986 May). https://doi.org/10.1037/0033-2909.99.3.432.

[51] M. Caetano, C. Saitis, and K. Siedenburg, "Audio Content Descriptors of Timbre," in K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, and R. R. Fay (Eds.), *Timbre: Acoustics, Perception, and Cognition*, Springer Handbook of Auditory Research, vol. 69, pp. 297–333 (Springer, Cham, Switzerland, 2019). https://doi.org/10.1007/978-3-030-14832-4_11.

[52] G. Peeters, "The Timbre Toolbox: Extracting Audio Descriptors from Musical Signals," *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 2902–2916 (2011 Nov.).

[53] J. Štěpánek, "Musical Sound Timbre: Verbal Description and Dimensions," in *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx)*, pp. 121–126 (Montreal, Canada) (2006 Sep.).

[54] J. Noble, E. Thoret, M. Henry, and S. McAdams, "Semantic Dimensions of Sound Mass Music: Mappings Between Perceptual and Acoustic Domains," *Music Percept.*, vol. 38, no. 2, pp. 214–242 (2020 Nov.). https://doi.org/10.1525/mp.2020.38.2.214.

[55] C. Saitis and K. Siedenburg, "Brightness Perception for Musical Instrument Sounds: Relation to Timbre Dissimilarity and Source-Cause Categories," *J. Acoust. Soc. Am.*, vol. 148, no. 4, pp. 2256–2266 (2020 Oct.). https://doi.org/10.1121/10.0002275.

[56] M. J. Vowels and R. Mason, "Comparison of Pairwise Dissimilarity and Projective Mapping Tasks With Auditory Stimuli," *J. Audio Eng. Soc.*, vol. 68, no. 9, pp. 638–648 (2020 Sep.). https://doi.org/10.17743/jaes.2020.0051.

[57] S. S. Stevens, "Tonal Density," *J. Exp. Psychol.*, vol. 17, no. 4, pp. 585–592 (1934 Aug.).

[58] E. J. Allen and A. J. Oxenham, "Symmetric Interactions and Interference Between Pitch and Timbre," *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1371–1379 (2014 Mar.). https://doi.org/10.1121/1.4863269.

[59] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of Valence, Arousal, and Dominance for 13,915 English Lemmas," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1191–1207 (2013 Feb.). https://doi.org/10.3758/s13428-012-0314-x.

[60] R. Vertegaal and E. Bonis, "ISEE: An Intuitive Sound Editing Environment," *Comput. Music J.*, vol. 18, no. 2, pp. 21–29 (1994 Summer). https://doi.org/10.2307/3680440.

[61] J. R. Lam and C. Saitis, "The Timbre Explorer: A Synthesizer Interface for Educational Purposes and Perceptual Studies," in *Proceedings of the New Interfaces for Musical Expression Conference (NIME)* (Shanghai, China), paper 62 (2021 Jun.). https://doi.org/10.21428/92fbeb44.92a95683.

[62] A. Zacharakis, K. Pastiadis, and J. D. Reiss, "An Interlanguage Unification of Musical Timbre: Bridging Semantic, Perceptual, and Acoustic Dimensions," *Music Percept.*, vol. 32, no. 4, pp. 394–412 (2015 Apr.). https://doi.org/10.1525/mp.2015.32.4.394.

[63] M. Huzaifah and L. Wyse, "Deep Generative Models for Musical Audio Synthesis," in E. R. Miranda (Ed.), *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*, pp. 639–678 (Springer, Cham, Switzerland, 2021).

[64] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," presented at the *8th International Conference on Learning Representations* (Addis Ababa, Ethiopia) (2020 Apr.).

[65] S. Huang, Q. Li, C. Anil, et al., "TimbreTron A WaveNet(CycleGAN(CQT(Audio))) Pipeline for Musical Timbre Transfer," presented at the *7th International Conference on Learning Representations* (New Orleans, LA) (2019 May).

[66] B. Hayes, C. Saitis, and G. Fazekas, "Neural Waveshaping Synthesis," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, pp. 254–261 (2021 Nov.).

[67] G. Bromham, D. Moffat, M. Barthet, A. Danielsen, and G. Fazekas, "The Impact of Audio Effects Processing on the Perception of Brightness and Warmth," in *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*, pp. 183–190 (Nottingham, UK) (2019 Sep.). https://doi.org/10.1145/3356590.3356618.

A.1 TOP 50 TIMBRE DESCRIPTIONS

The 50 most frequently used timbral adjectives collected from a popular modular synthesis forum according to the procedure described in SEC. 1.2.

Table 6.

	Description	Bigram Occ.	Corpus Occ.
1	Great	12,637	128,040
2	Good	6,158	142,535
3	Nice	3,584	92,787
4	Different	3,271	80,763
5	Awesome	1,896	32,652
6	Cool	1,734	54,245
7	Amazing	1,571	20,479
8	Interesting	1,415	40,124
9	Fantastic	1,286	9,598
10	Synth	1,222	60,582
11	Percussive	1,217	3,482
12	Pretty	1,093	75,287
13	Similar	1089	29,786
14	New	887	88,297
15	Unique	848	8,253
16	Beautiful	692	9,237
17	Digital	678	30,144
18	Clean	670	12,526
19	Complex	573	15,652
20	Incredible	555	4,106
21	Modular	552	118,712
22	FM	540	27,389
23	Wonderful	536	6,525
24	Overall	516	5,425
25	Right	491	77,903
26	Bad	487	23,048
27	Weird	446	12,432
28	Excellent	446	11,666
29	Drum	437	41,217
30	Organic	419	2,383
31	Sweet	409	8,992
32	Crazy	408	11,627
33	Raw	385	3,557
34	External	372	22,864
35	Natural	364	2,684
36	Fine	362	32,489
37	Basic	352	19,560
38	Classic	345	8,470
39	Original	330	23,436
40	Electronic	323	9,695
41	Much	322	120,812
42	Many	315	56,465
43	Huge	307	11,131
44	Rich	302	3,003
45	Big	300	34,466
46	Metallic	297	1,268
47	Musical	296	9,838
48	Specific	293	12,207
49	Decent	288	9,692
50	Certain	279	11,501

A.2 DESCRIPTOR PRUNING CRITERIA

1. Remove words referring to affect (e.g., *good*).
2. Remove words referring to specific synthesizers or hardware (e.g., *moogy*).
3. Keep only one element of any group of words sharing a stem, favoring the word with the highest corpus frequency (e.g., *wooden* and *woody*).
4. Remove words more commonly used to describe pitch than timbre (e.g., *high*).
5. Remove words describing loudness (e.g., *loud*).
6. Remove words describing duration (e.g., *short* or *long*).
7. Keep only one element of any group of obvious synonyms (e.g., *brilliant* and *bright*).

A.3 EXPLANATION OF COMPARATIVE FACTOR MODEL

Let X be the full matrix of unobserved absolute semantic ratings. Let X_c and X_r be matrices such that the sets of rows of X_c and X_r are overlapping subsets of the set of rows of X, with X_c containing ratings of sounds created by participants and X_r containing ratings of the reference sounds.

The theoretical factor model $X = LF + M + \varepsilon$, where F is the matrix of factor scores for each observation and each column of matrix M contains the mean of the corresponding column of X and then gives the overall loading matrix L with which models for X_c and X_r, $X_c = LF_c + M_c + \varepsilon_c$ and $X_r = LF_r + M_r + \varepsilon_r$, can be specified. This loading matrix thus also applies to the model of observed comparative ratings:

$$\begin{aligned} X_{diff} &= X_c - X_r + \epsilon_{diff} \\ &= L(F_c - F_r) + M_c - M_r + \epsilon_c - \epsilon_r + \epsilon_{diff} \\ &= LF_{diff} + M_{diff} + \epsilon. \end{aligned}$$

Again, by linearity, the difference in the column means (M_c and M_r) of X_c and X_r is equal to the column mean of the element-wise differences between X_c and X_r, giving M_{diff}. The respective error terms (ε_c and ε_r) of these implicit absolute models are, on account of their normality, simply subsumed into the error term of the observed comparative model as a sum of normally distributed random variables.

THE AUTHORS


Ben Hayes


Charalampos Saitis


György Fazekas

Ben Hayes is a Ph.D. student at the Centre for Digital Music (C4DM) in the School of Electronic Engineering and Computer Science at Queen Mary University of London (QMUL), United Kingdom, where he works under the supervision of Charalampos Saitis and György Fazekas as part of the UK Research and Innovation Centre for Doctoral Training in Artificial Intelligence and Music. His research centers around novel applications of deep learning for modeling the synthesis and perception of musical timbre, with a particular focus on meta-learning techniques. He also holds an M.Sc. degree in Sound and Music Computing from QMUL and B.Mus.(Hons.) in Electronic Music from the Guildhall School of Music and Drama. He is an organizing member of the Special Interest Group on Neural Audio Synthesis at C4DM, and in December 2021 he organized the first international Neural Audio Synthesis Hackathon. Previously he worked as music lead at generative music startup Jukedeck, where he contributed to their successful acquisition by ByteDance. He has also toured internationally as a musician and is currently signed to R&S Records.

•

Charalampos Saitis studied Mathematics and Musical Acoustics in Athens and Belfast and obtained a Ph.D. in Music Technology from McGill University. He is currently Lecturer at the Centre for Digital Music of Queen Mary University of London and Turing Fellow at the Alan Turing Institute. His research concerns communication acoustics with a focus on timbre perception, sensory cross-modality, and "metaphors we listen with." He acted as co-editor of the Springer Series on Touch and Haptic Systems volume *Musical Haptics* (2018) and the Springer Handbook of Auditory Research volume *Timbre: Acoustics, Perception, and Cognition* (2019), and he has authored several recent publications on timbre perception and semantics. He was co-organizer of the Berlin Interdisciplinary Workshop on Timbre (2017) and a founding member of the International Conference on Timbre (2020).

•

György Fazekas is a Senior Lecturer at the Center for Digital Music, Queen Mary University of London (QMUL). He holds a B.Sc., M.Sc. and Ph.D. degree in Electronic Engineering. He is an investigator of UK Research and Innovation's £6.5M Centre for Doctoral Training in Artificial Intelligence and Music, and he was QMUL's Principal Investigator on the H2020 funded Audio Commons project. He was general chair of Association for Computing Machinery's Audio Mostly 2017 and papers co-chair of the AES 53rd International Conference on Semantic Audio, and he received the Citation Award of the AES. He published over 150 papers in the fields of music information retrieval, semantic web, deep learning, and semantic audio.