Disembodied Timbres: a Study on Semantically Prompted FM Synthesis

Ben Hayes, Charalampos Saitis, György Fazekas
{b.j.hayes, c.saitis, 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 generalise to electronic sounds, nor is it obvious how these relate to the creation of such sounds. In this work, we present 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. We applied a novel experimental paradigm in which experienced sound designers responded to semantic prompts by programming a synthesiser, and provided semantic ratings on the sounds they created. 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. Our results suggest that further enquiry 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, as well as synthesis control and audio engineering.

1 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 our auditory experience. It enables us to identify who is speaking to us, to ascertain the source of a sound, and is of central importance to the aesthetic experience of music [1]. Increasingly, our timbral world is populated by sounds with no discernible physical source, which we refer to as *disembodied sounds*. Contemporary sound design tools and sound reproduction apparatus pair to enable us 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 we receive from movies and games, and have defined entirely new audio cultures [2]. Our understanding of timbre, however, is largely limited to insights gleaned from studies on musical instrument sounds playing isolated notes. In this work, we set 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 synthesise electronic sounds prompted with well established semantic dimensions of timbre.

Studying the perception of disembodied electronic sounds may help further elucidate the mechanisms underpinning our 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 a frequency modulation (FM) synthesiser. We also set out to identify whether prompting synthesis with semantic descriptors would result in a discernable impact on the control of synthesiser 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 synthesisers do not represent a significant progression from the approach of early synthesisers – 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 dis-

similarity 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 our perception. To facilitate further research in this direction, we make available the dataset of sounds generated in our study[1]. Alongside rendered audio of all synthesised sounds, we provide full parameter configurations, semantic ratings, acoustic features, and anonymous participant questionnaire responses.

1.1 From Sounds to Adjectives

The perception of timbre has enjoyed an extensive lineage of scientific enquiry, 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 [15]. 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 [16, 17, 18, 19]. 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 which involved learning kernel distance functions over data from 17 previous dissimilarity studies [20]. 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 generalising 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 crossmodal, 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 (EFA) 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 [21]) or by an adjective and its negation as in "rough–not rough" (known as the verbal attribute magnitude estimation method [22]).

This approach has a long history in empirical research on timbre, being first used in 1958 to study sonar sounds [23], about a decade before the early MDS studies of the 1970s [16, 17]. It was first applied to musical sounds by von Bismarck [24] in 1974, who used synthetic recreations of instrumental and vocal timbres. It has since been employed in numerous studies of musical timbre [25, 22, 26, 19, 27] (for a comprehensive review, see [7]). Despite differences in methodology (choice of verbal scales, 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. [27] performed an interlanguage study with musically experienced Greek- and English-speaking listeners, where responses from both linguistic groups were well explained by a model which 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 [28] 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 anaylsis 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 [29]. 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 potentially correspond to the three LTM factors [27]; "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. Where electronic and synthesised sounds do find use, it is typically either for the purposes of simulating the sounds of familiar acoustic instruments and the human voice [24, 25] or for the creation of controlled stimuli designed to elicit a specific perceptual response [30, 31]. It is not currently clear how well these multidimensional semantic models might generalise 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

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

texture might generalise beyond the musical instrument domain. However, the tested textural sounds involved multiple different timbres and/or 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.

1.2 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 [32, 33, 34, 35, 36, 37]. To explore this question, here we used a semantically prompted FM synthesis task and examined semantic associations of timbre 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 [38] 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 synthesisers, 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 synthesiser 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 synthesisers, 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. [18] used simulations of traditional western instruments synthesised by Wessel et al. [39] on a Yamaha TX802 FM Tone Generator. An earlier study of timbre semantics by Ashley [40] 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 to attract a varied aesthetic vocabulary, whilst being abstract enough to avoid inherently implying a distinct source-cause. Wallmark et al. [41] 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 two-dimensional space that linearly mapped to the controls of a simple FM synthesiser 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.

1.3 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 [41] 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 synthesiser enthusiasts, and the study focused on timbres created by music and audio technologists with experience in sound design and synthesis. We carried out exploratory factor analysis of comparative semantic ratings and principal components analysis of acoustic features extracted from the created sounds. Linear regression and correlation analyses subsequently enabled us to quantify the interrelations between language, psychoacoustics, and the adjustment of synthesiser controls.

Where the design of Wallmark et al. [41] focused solely on the effects of spectral energy distribution, as participants were shaping only static aspects of a continuous tone, here we sought to incorporate spectrotemporal and purely temporal aspects of the FM sounds by providing a full set of amplitude envelope controls. We also applied three distinct fundamental frequency (F0) conditions for each comparative prompt. In research on timbre it is usual to equalise the F0 of stimuli as pitch and timbre are known to interact [42, 43]. Here we wanted to explore whether such interaction would exert an effect on synthesiser parameter control, that is, on shaping timbre. We also wanted to examine the influence of F0 on the semantic dimensions of FM sounds.

2 METHOD

2.1 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 (GoldMSI) inventory [44]. Compared to the reference statistics provided with GoldMSI, 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.

2.2 Word Stimuli

To maximise the appropriateness of word stimuli selection to synthesised sounds we adopted a corpus-based approach, mining descriptors from a popular modular synthesis forum[2]. We collected publicly available posts from the forum dating up to 21st February 2020, for a total of 1,407,604 posts. After lemmatisation, 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 NLTK'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 [27]. For example, *brilliant* and *bright* loaded highly onto the *luminance* factor. We then retained the word with the highest frequency in our corpus for each factor – e.g. *bright* in the case of the luminance factor.

2.3 Synthesiser

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) [38]. This is illustrated by equation 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 equation 1 strictly describes *phase modulation* rather than frequency modulation, which produces an equivalent magnitude spectrum when using sinusoidal oscillators. As FM synthesisers are typically implemented with phase modulation, we used this formulation for our experimental synthesiser.

The synthesiser 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 ADSR (Attack, Decay, Sustain, Release) 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. Our experimental synthesiser is thus given by equation 2:

$$x(t) = A\varepsilon_1(t)\sin(\omega_1 t + I_2\varepsilon_2(t)\sin\omega_2 t + I_3\varepsilon_3(t)\sin\omega_3 t), \tag{2}$$

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 synthesisers, the operator tuning ratio parameters were divided across two controls:

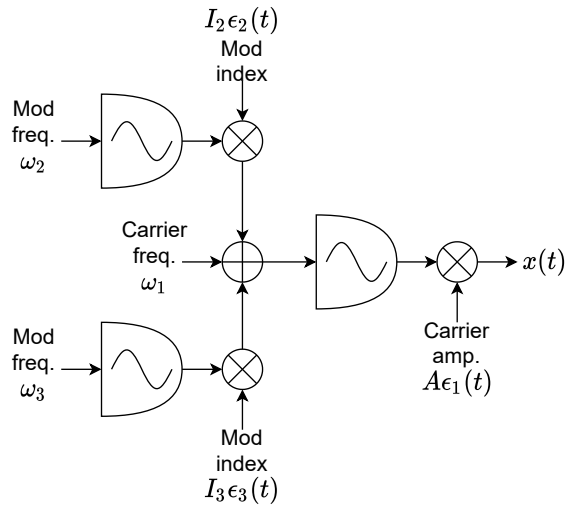

Fig. 1. A schematic diagram of our three operator frequency modulation synthesiser

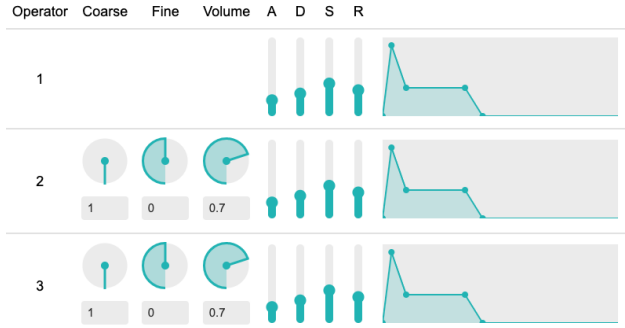

Fig. 2. The FM synthesis interface used by the participants

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

coarse and *fine*. The *coarse* control specified the integer part of the tuning ratio, whilst the *fine* control specified the 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, as 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.

2.4 Procedure

Due to 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 [45]. The study was built using *lab.js* [46] and the WebAudio API's AudioWorklet was used to build a real-time in-browser FM synthesiser.[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*, *rougher* or *less rough*) and three pitches (E2, A3, 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 synthesiser 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 parameters p_c to fulfil the given comparative prompt (e.g., to create a sound that is *brighter* or *less bright* than the reference).

2(a) 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 a resolution of 0.1.

2(b) 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 a 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 synthesiser'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 synthesiser were initialised identically for all participants. Given the sound design expertise of the participants, this approach also enabled us to focus our analysis on regions of synthesiser parameter space that are of interest to experienced synthesists. Limitations of this approach are discussed in section 4.3. To start this process, the database was initialised with a starting set of nine "seed" sounds, which were hand-designed by the first author and loosely based on popular DX7 patches.

3 RESULTS

3.1 Exploratory Factor Analysis

We conducted initial reliability analyses using Cronbach's α. All 27 semantic scales showed high internal consistency, average $\alpha = .95$ and $\sigma = .003$. 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, we assume 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. [47] provide a review of such methods and a discussion of their strengths and weaknesses. Amongst the most popular are the Kaiser criterion, Cattell's scree test, and Horn's parallel analysis. The Kaiser criterion [48] involves retaining as many factors as there are eigenvalues of the correlation matrix ≥ 1.0. In Cattell's [49] method, a scree plot (correlation matrix eigenvalues plotted against their indices) is inspected with the aim of identifying an "elbow" point which signifies an appropriate number of factors. Horn's parallel analysis [50] is a bootstrap method in which an identical factor analysis procedure is conducted on a large number of normally distributed random datasets of

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

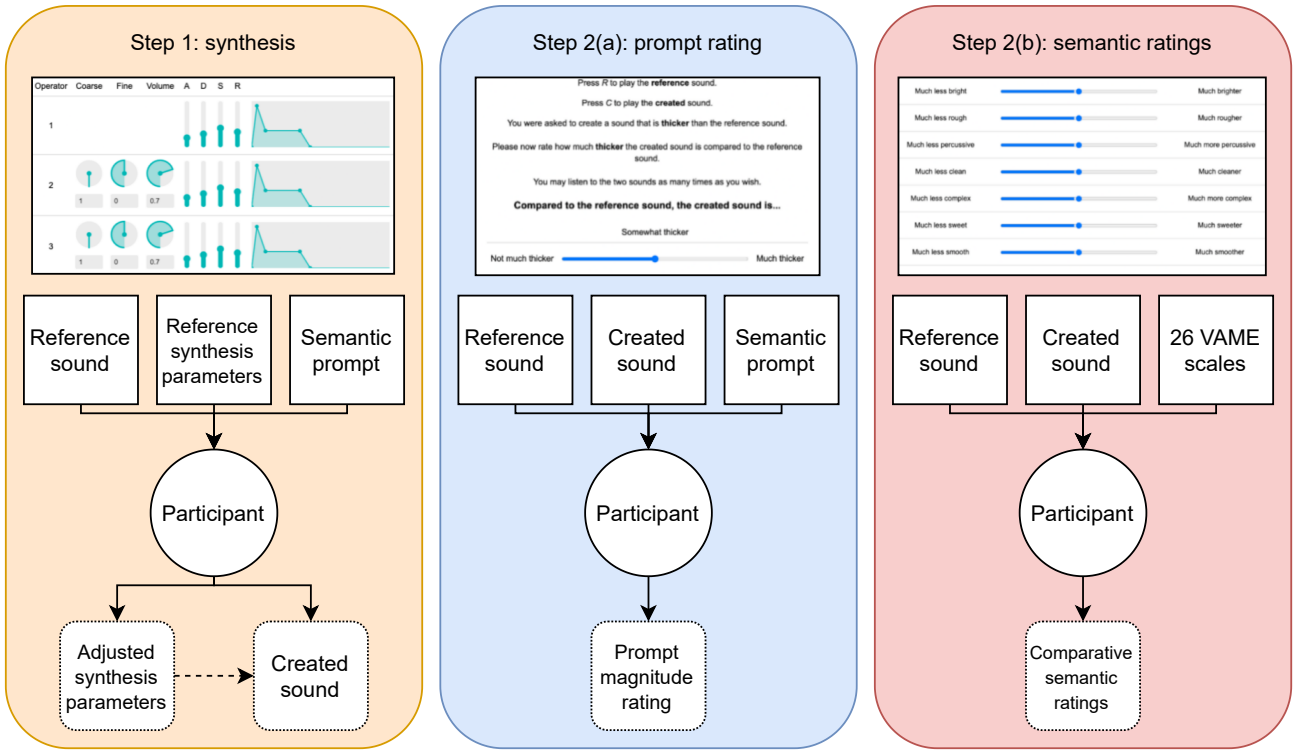

Fig. 3. A schematic diagram illustrating the experimental procedure for a single trial, repeated for each prompt and register. Step 1 (orange): participants synthesise a sound in response to a prompt. Step 2(a) (blue): participants rate the difference between the reference sound and their created sound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

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 [47] and synthetic [51] 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 we explored a semantic space for the created timbres 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 due to 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*, *harsh*). Factor F2 showed strong loadings onto terms related to mass (*big*, *thick*, and negatively *thin*).

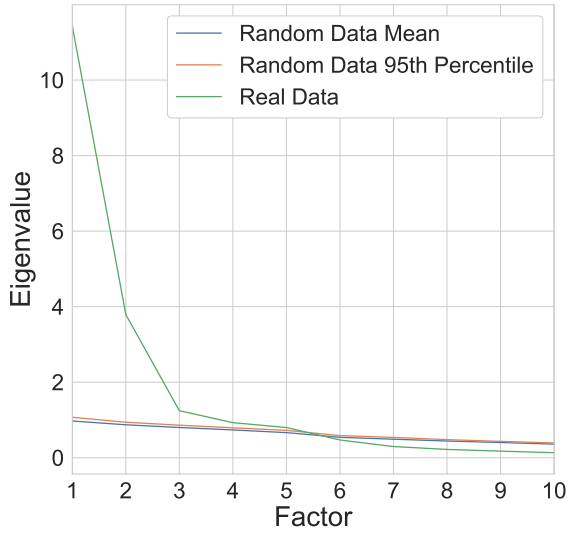

Fig. 4. A scree plot comparing the factor eigenvalues of our dataset to the mean and 95th percentile of the factor eigenvalues of the stochastic datasets generated in parallel analysis. Here, we see that the procedure supports 5 factors at the 95th percentile level.

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

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	**.82**	−.07	.06	.16	.07
metallic	**.75**	.05	−.05	.09	.11
bright	**.73**	−.22	.04	.10	.05
harsh	**.72**	.01	−.18	.08	.15
big	.30	**.87**	−.03	−.16	−.04
thick	−.15	**.84**	−.10	.02	−.04
deep	−.43	**.70**	.00	−.07	.06
thin	.32	**−.70**	.20	.11	.02
clean	−.04	.02	**.90**	−.02	−.01
clear	.17	−.04	**.78**	.07	−.03
plucky	−.04	−.09	.07	**.99**	−.05
percussive	.04	−.02	−.06	**.78**	.06
raw	.01	−.12	.12	.01	**.78**
rich	.32	.69	.08	−.06	−.03
dull	−.69	−.12	.02	−.25	−.03
mellow	−.67	−.04	.17	−.12	−.15
woody	−.63	.20	.01	.23	−.18
warm	−.60	.42	.17	−.06	−.01
dark	−.58	.51	.06	−.05	.19
aggressive	.57	.27	−.06	.15	.33
sweet	−.03	.13	.43	.10	−.56
noisy	.52	.10	−.40	.11	.12
hard	.49	.24	−.14	.24	.23
smooth	−.49	.00	.40	−.24	−.08
complex	.48	.36	−.35	.10	−.11
gritty	.48	.26	−.32	.18	.17
rough	.42	.16	−.26	.21	.29

Bold type indicates loadings with an absolute value greater than .70.

basis of either the highest-loading word (F1 and F5) or one that we thought would better capture the meaning of the corresponding dimension (F2–F4).

Table 2. Inter-factor correlations and angles

	F1	F2	F3	F4
F2	−.08 *(94.4°)*			
F3	−.42 *(114.6°)*	−.30 *(107.7°)*		
F4	.51 *(59.3°)*	−.17 *(99.6°)*	−.27 *(105.4°)*	
F5	.37 *(68.3°)*	.07 *(85.8°)*	−.44 *(116.2°)*	.31 *(72.1°)*

Table 2 reports the inter-factor correlation coefficients (r) after rotation, as well as 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 odbserved with F2, suggesting that impressions of *mass* in these FM sounds might have been perceptually more distinct from the other four semantic dimensions.

3.2 Acoustic Features Analysis

To study the psychoacoustic underpinnings of the semantic space, a large set of acoustic features were extracted from the created sounds. Spectral features were computed on multiple representations, namely STFT magnitude and

power spectra, Bark frequency magnitude spectrum, and harmonic peak magnitudes [52]. Further, harmonic features including inharmonicity, odd-to-even ratio, and tristimulus, and purely temporal features including log attack time, temporal centroid, and zero-crossing rate were computed.

Spectral features were computed using a Hamming window of size 1024 with an overlap of 75%, and silent frames were discarded. Framewise features were summarised by the median and interquartile range. All features were computed using the Essentia library for Python. Synthesiser patches were rendered at 44.1kHz with a duration of 4 seconds. The ADSR envelope was controlled by a gate signal which was on (attack, decay, and sustain) for 3 seconds, and off (release) for 1 second.

The extracted features can not 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, we followed a feature dimensionality reduction procedure based on that of Zacharakis et al. [27]. 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.

Whilst 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. We found that 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, we introduced an extra step before the reduction procedure wherein features were sorted by their maximum absolute Spearman rank correlation coefficent with any of the semantic factors. Then, the member of each collinear feature pair with the lowest such factor correlation was discarded. We believe this filter-based approach to be sufficient for the task of identifying acoustical correlates and thus leave deeper analysis of features and the application of alternate feature selection methods to future work.

Owing to the large number of features computed, we set our threshold for the Spearman rank correlation coefficient at 0.7 and for the KMO measure of sampling adequecy at 0.7. This resulted in a set of 17 descriptors, which are listed in Table 3. Parallel analysis, performed on the resulting set of features, supported a 4 component solution at the 95th percentile level. PCA was followed by Varimax rotation to achieve simple structure. The resulting component loadings are shown in Table 3. Features with loadings above a threshold (set at 0.75) are used to label components.

The first component shows above-threshold loadings for the medians of spectral decrease [53], 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 for 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 synthesiser'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 4 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 ($n = 270$), as well as the inherent noise in the dataset caused by the single rating provided for each sound and the subjectivity of assigning a value to the applicability of a semantic descriptor. In particular, whilst 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.

Table 3. Principal component loadings of acoustic features after varimax rotation

	PC1	*PC2*	*PC3*	*PC4*
	Spectrotemporal (distribution) & spectral shape	*Temporal energy variation & spectral slope*	*Spectrotemporal (flatness)*	*Spectrotemporal (crest factor)*
$STFT_{pow}$ kurtosis IQR	**1.00**	.00	.00	.00
$STFT_{pow}$ skewness IQR	**.95**	.08	−.30	.02
Bark spread median	**.82**	.57	−.02	−.11
$STFT_{mag}$ decrease median	**.78**	.48	−.40	.02
Bark crest median	**.76**	.61	.23	−.03
Harmonic kurtosis IQR	**.76**	−.13	−.19	−.61
$STFT_{pow}$ frame erg IQR	.00	**1.00**	.00	.00
Harmonic frame erg IQR	.46	**.82**	.20	−.28
Effective duration	−.44	**.80**	.36	−.21
Harmonic decrease median	.10	**.76**	.16	.62
$STFT_{mag}$ flatness IQR	.00	.00	**1.00**	.00
$STFT_{pow}$ crest IQR	.19	−.21	**.94**	.19
$STFT_{mag}$ crest IQR	.00	.00	.00	**1.00**
$STFT_{mag}$ centroid IQR	.67	.31	−.47	.48
$STFT_{pow}$ kurtosis median	.69	.42	.13	.58
$STFT_{pow}$ skewness median	−.18	.71	−.64	−.22
Bark centroid median	.66	.72	.09	.18

Bold type signifies absolute component loading > 0.75. Features with loadings at this level are used to label components, as in [27]

3.3 Synthesiser Parameters

We next set to inspect the perceptual imprints of timbre on the sound design process. In order to identify whether semantic prompts and the direction of comparison exerted an effect on the adjustments made to synthesiser 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).

We observed similar patterns of linear effects on changes to the modulator tuning and volume parameters for *brighter*, *less bright*, and *less rough* prompts, with the polarity of the effects inverted for the "less" prompts. These effects were also present for *rougher*, though 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 the 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 response to prompts. In the case of modulator controls, this may be explained by their equivalence in the architecture of our synthesiser – 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, we observed that prompt effects on tuning and volume controls were consistently stronger than on ADSR envelope controls. This may be partially due to 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 due mostly to the lack of a prompt that explicitly describes temporal characteristics of a signal. As an explicit percussive-plucky factor emerged in our 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 synthesiser controls, semantic factors, and the principal axes of acoustical variation, we computed Spearman's rank correlation coefficient computed between synthesiser 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. Corre-

Table 4. 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)*	$-.58^{***}$	$-.37^{***}$	$.49^{***}$	$-.25^{***}$	$-.01$
Factor 2 *(Mass)*	.09	$-.02$	.09	.03	.08
Factor 3 *(Clarity)*	$.29^{***}$	$.17^{**}$	$-.44^{***}$	.04	$-.03$
Factor 4 *(Percussiveness)*	$-.24^{***}$	$-.03$	$.31^{***}$	$-.14^{*}$	$-.02$
Factor 5 *(Rawness)*	$-.22^{***}$	$-.10$	$.34^{***}$	$-.10$	$-.05$

$^{*}: p < 0.05$; $^{**}: p < 0.01$; $^{***}: p < 0.001$

lations 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 synthesised sounds. Modulator volume, however, appeared to exhibit almost no relationship with factor F2 (*mass*), whilst 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 enevelope 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 characterised 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 due to the lack of a specifically temporal prompt descriptor.

4 DISCUSSION

We explored the semantic correspondences of a wide variety of sounds produced through FM synthesis 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. We studied these responses by constructing a semantic timbre space using exploratory factor analysis and performed a correlation analysis with the principal components of a set of acoustic features. Finally, we examined the influence of semantic prompts on the sound design process by fitting linear models to synthesiser 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 [27, 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 generalise well across timbral domains, whilst 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 can not be assumed to generalise beyond the timbral domain of our experimental FM synthesiser. Continued enquiry into the full diversity of electronic sound is needed to understand the extent to which our findings are due to specificities of FM synthesis.

4.1 Implications for the perception and semantic processing of timbre

The first factor, which we labelled *sharpness*, showed strong loadings for both luminance and texture related words, though 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* [54]. 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 [27] 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 our 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 synthesiser 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 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 between brightness and roughness which could not be entirely captured by our factor model.

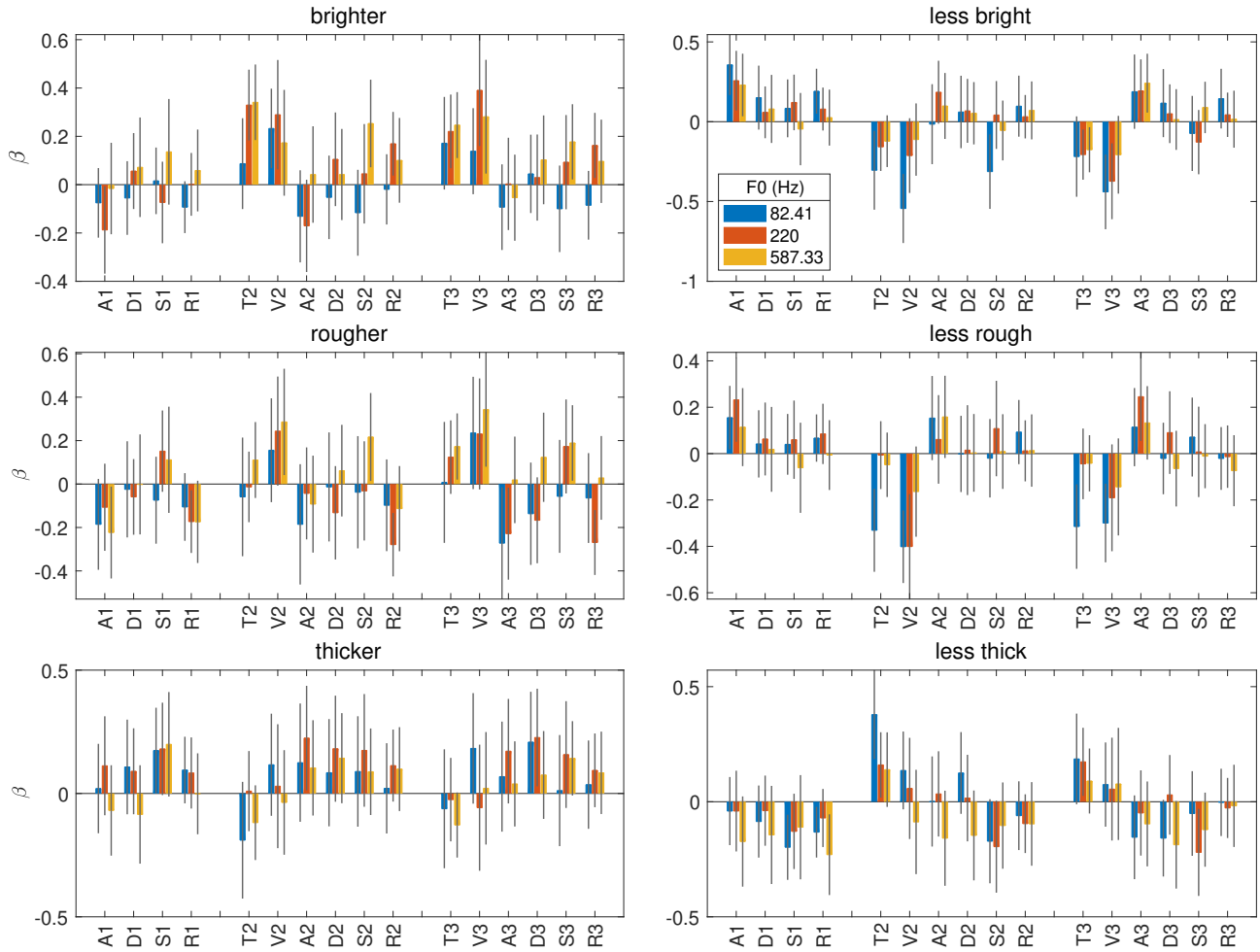

Fig. 5. Linear effects (β) of comparative semantic prompt derived from linear regression for every FM synthesiser control change and fundamental frequency. Error bars correspond to 95% confidence intervals. A = attack; D = decay; S = sustain; R = release; T = tuning; V = volume; 1 = carrier; 2/3 = modulators.

Acoustic correlations for the second semantic factor (*mass*) were less clear. On the one hand, this might be the result of our 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 our dataset has obscured any such individual correlations. Indeed, the two most highly correlated acoustic principal components (PC1 & 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 synthesiser sounds are best characterised 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 [55].

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 synthesised 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 [38]. It is plausible that certain aspects of variation between FM synthesised 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 due to high correlation with other scales [27].

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 synthesiser'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, 28].

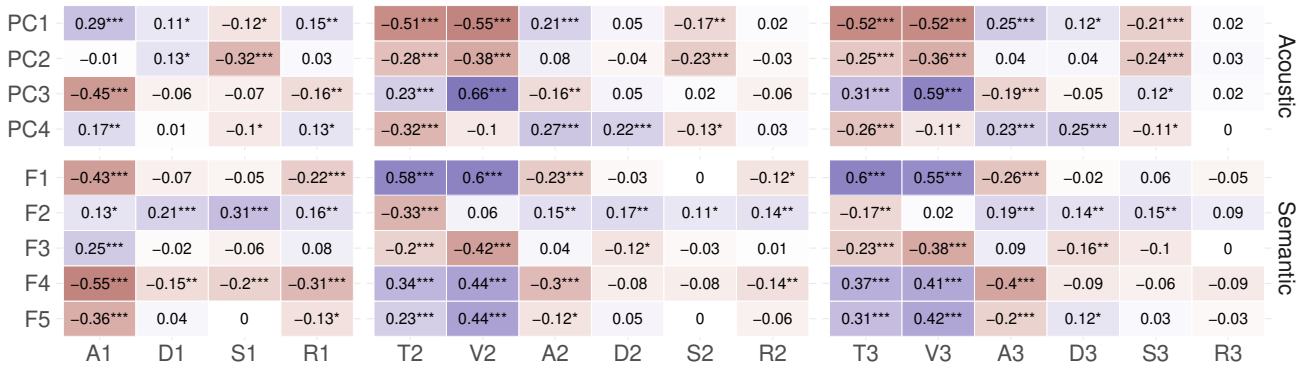

Fig. 6. Spearman's ρ computed between changes to synthesiser controls, semantic factors, and differences along acoustic principal components. F1-F5 = semantic factors; PC1-PC4 = acoustic prinicpal components; A = attack; D = decay; S = sustain; R = release; T = tuning; V = volume; 1 = carrier; 2 & 3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

Whilst 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 synthesised sounds, the attributes insinuated by the terms *percussive* and *plucky* are not well characterised 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 our analysis and previous work [19, 27] as, 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.

4.2 Relationship between semantic factors and synthesis parameters

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

1. **Amplitude temporal evolution**: carrier attack (A1), decay (D1), sustain (S1), release (R1).
2. **Spacing between sideband frequencies**: modulator tuning (T2, T3)
3. **Sideband energy distribution**: modulator volume (V2, V3)
4. **Sideband energy temporal evolution**: modulator attack (A2, A3), decay (D2, D3), sustain (S2, S3), release (R2, R3).

With these groupings in mind, analysing 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, ↓R1), (2) wider spacing between sidebands (↑T2, ↑T3), (3) more energy distributed to sidebands (↑V2, ↑V3), and (4) a shorter sideband energy envelope (↓A2, ↓A3). Conversely, increasing "mass" (F2) suggests parameter changes that cause (1) slower amplitude envelopes with more sustain (↑D1, ↑S1, ↑R1), (2) narrower spacing between sidebands (↓T2, ↓T3) (3) no change to sideband energy distribution, and (4) slower sideband energy envelopes with more sustain (↑A2, ↑D2, ↑R2, ↑A3, ↑D3, ↑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 [28]. 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.

4.3 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, [56]). As such, the latter may have invited for subtler semantic associations. Reusing previously created sounds as reference stimuli for each trial may also have 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, as 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 not to limit the dimensionality that can be recovered [56]. As 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 which may be obscured in an absolute rating paradigm might have been enabled (although see [57]). 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 [42, 43] and when considered as an additional variable [27]. 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 (c.f. [58]). 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 [59].

The architecture of the FM synthesiser used by participants may have limited the power of the linear models presented in Section 3.3 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, whilst 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 our results derive from the nonlinearity and complexity of FM synthesis.

Further, a given semantic prompt may not map uniquely to a single point in the synthesiser's parameter space as per the instructed task. This is due to both the previously discussed symmetry of the synthesiser and to the fact that the synthesiser's parameters may not map directly onto the semantic dimensions under test. For example, it is plausible that the neighbourhood surrounding a "bright" sound in the parameter space also consists largely of "bright" sounds. It is also conceivable that there may exist several disjoint neighbourhoods 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 synthesiser's parameter space, as they provide only point estimates. Further research aiming to map these distributions across the ranges of parameters would therefore be valuable.

4.4 Influence of word affect on timbre-semantic associations

In the prompted synthesis study of Wallmark et al. [41], affective connotations of the adjective prompts (based on validated affect norms [60]) 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. We observed largely similar trends 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*) [60]. 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 3 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.

4.5 Towards perceptually-informed sound design and synthesis

As observed in the present study, and in previous work [8], the controls of existing synthesisers 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 synthesisers.

Wessel [9] first suggested the use of a timbre dissimilarity space, constructed using multidimensional scaling, as a control space for a synthesiser. 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 which oscillates between harmonic structure and total inharmonicity. Thus, simple timbre space mappings to FM parameters can be more challenging to derive [61, 62]. Further, mapping synthesis parameters to a semantic timbre space introduces yet another layer of complexity as, whilst timbre-semantic dimensions are assumed to relate to an underlying perceptual representation, the nature of this relationship is not clear for all dimensions [19, 63].

As research in neural audio synthesis [64] extends the capabilities of synthesisers 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 modelling [65], transfer the timbre of one instrument to another [66], perform perceptually smooth "morphs" between timbres [10], and more. Recent work [67] has enabled many of these techniques to be achieved comfortably in real time on consumer CPUs, 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 learnt 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 synthesisers 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 equalisation, compression, reverb [35], distortion [36], and bit-depth reduction [68]. 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, we accompany this work with a fully annotated dataset[3] of sounds produced in our study, with complete semantic ratings and factor loadings. We intend this as a first step towards sharing insights across these fields in a manner that will facilitate progress on this problem.

5 CONCLUSIONS

In this study we investigated the semantic associations of disembodied electronic timbres – specifically, those produced by a three operator FM synthesiser. We applied a novel experimental paradigm in which participants directly synthesised sounds in response to semantic prompts linked to the dimensions of the luminance-texture-mass model of timbre semantics. An exploratory factor analysis of comparative semantic ratings collected between pairs of synthesised sounds recovered a five factor semantic space. To identify the acoustic underpinnings of the resulting factors, we performed a correlation analysis with the principal components of a comprehensive set of acoustic features. We also fit linear regression models to examine the effects of semantic prompts on the use of synthesiser controls.

Semantic factors corresponding to *luminance*, *texture*, and *mass* (LTM) were present in our model, but *luminance* and *texture* were combined. We found acoustic correlates of *luminance* and *texture* similar to those observed in previous work [27], 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 synthesiser 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 our model's three novel dimensions would be valuable. The application of classical timbre dissimilarity and semantics paradigms to sounds generated in our study would also facilitate interpretation of these results in the broader context of timbre research.

6 ACKNOWLEDGMENTS

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.

7 REFERENCES

[1] K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, R. R. Fay (Eds.), *Timbre: Acoustics, Perception, and Cognition* (Springer, Cham) (2019 May).

[2] C. Fales, "Hearing timbre: Implicit perceptual learning among early Bay Area ravers," in R. Fink, M. Latour, Z. Wallmark (Eds.), *The Relentless Pursuit of Tone: Timbre in Popular Music*, pp. 21–42 (Oxford University Press, New York, NY) (2018 Oct.), doi: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, pp. 608–619 (2018 Aug.), doi:10.1121/1.5048130.

[4] C. Vahidi, G. Fazekas, C. Saitis, A. Palladini, "Timbre Space Representation of a Substractive Synthesiser," presented at the *Proceedings of the 2nd International Conference on Timbre* (2020 Sep.).

[5] T. Grill, A. Flexer, S. Cunningham, "Identification of perceptual qualities in textural sounds using the repertory grid method," presented at the *Proceedings of the 6th Audio Mostly Conference on A Conference on Interaction with Sound - AM '11*, pp. 67–74 (2011 Sep.), doi:10.1145/2095667.2095677.

[6] M. Carron, T. Rotureau, F. Dubois, N. Misdariis, P. Susini, "Speaking about sounds: a tool for communication on sound features," *Journal of Design Research*, vol. 15, no. 2, pp. 85–109 (2017 Sep.), doi:10.1504/JDR.2017.086749.

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

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

[9] D. L. Wessel, "Timbre Space as a Musical Control Structure," *Computer Music Journal*, vol. 3, no. 2, p. 45 (1979 Jun.), doi:10.2307/3680283.

[10] P. Esling, A. Chemla, A. Bitton, "Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbre spaces," presented at the *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pp. 175–181 (2018 Sep.).

[11] S. Le Groux, P. F. Verschure, "Perceptsynth: mapping perceptual musical features to sound synthesis parameters," presented at the *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 125–128 (2008 Mar.), doi:10.1109/ICASSP.2008.4517562.

[12] S. Conan, E. Thoret, A. Mitsuko, O. Derrien, G. Charles, S. Ystad, R. Kronland-Martinet, "An Intuitive Synthesizer of Continuous Interaction Sounds: Rubbing, Scratching and Rolling," *Computer Music Journal*, vol. 38, no. 4 (2014 Dec.), doi:10.1162/COMJ_a_00266.

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

[14] H. Helmholtz, *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik* (F. Vieweg und Sohn, Braunschweig), 4th ed. (1877), english edition: H. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music* (Dover, New York), trans: A. J. Ellis, 2nd ed. (1954).

[15] K. Siedenburg, C. Saitis, S. McAdams, "The Present, Past, and Future of Timbre Research," in K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, R. R. Fay (Eds.), *Timbre: Acoustics, Perception, and Cognition*, pp. 1–19 (Springer International Publishing, Cham) (2019 May), doi:10.1007/978-3-030-14832-4_1.

[16] R. Plomp, "Timbre as a Multidimensional Attribute of Complex Tones," in *Frequency Analysis and Periodicity Detection in Hearing*, pp. 397–410 (1970).

[17] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *The Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277 (1977), doi:10.1121/1.381428.

[18] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychological Research*, vol. 58, no. 3, pp. 177–192 (1995 Dec.), doi:10.1007/BF00419633.

[19] T. M. Elliott, L. S. Hamilton, F. E. Theunissen, "Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. 389–404 (2013 Jan.), doi:10.1121/1.4770244.

[20] E. Thoret, B. Caramiaux, P. Depalle, S. McAdams, "Learning metrics on spectrotemporal modulations reveals the perception of musical instrument timbre," *Nature Human Behaviour* (2020 Nov.), doi:10.1038/s41562-020-00987-5, URL http://www.nature.com/articles/s41562-020-00987-5.

[21] C. E. Osgood, "The nature and measurement of meaning." *Psychological Bulletin*, vol. 49, no. 3, pp. 197–237 (1952), doi:10.1037/h0055737.

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

[23] L. N. Solomon, "Semantic Approach to the Perception of Complex Sounds," *The Journal of the Acoustical Society of America*, vol. 30, no. 5, pp. 421–425 (1958), doi:10.1121/1.1909632.

[24] G. von Bismarck, "Timbre of steady sounds: A factorial investigation of its verbal attributes," *Acustica*, vol. 30, pp. 146–159 (1974 Mar.).

[25] R. Pratt, P. Doak, "A subjective rating scale for timbre," *Journal of Sound and Vibration*, vol. 45, no. 3, pp. 317–328 (1976 Apr.), doi:10.1016/0022-460X(76)90391-6.

[26] A. C. Disley, D. M. Howard, A. D. Hunt, "Timbral description of musical instruments," presented at the *Proceedings of the 9th International Conference of Music Perception and Cognition and 6th Conference of the European Society for the Cognitive Sciences of Music* (2006 Aug.).

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

[28] A. Zacharakis, K. Pastiadis, "Revisiting the Luminance-Texture-Mass Model for Musical Timbre Semantics: A Confirmatory Approach and Perspectives of Extension," *Journal of the Audio Engineering*

Society, vol. 64, no. 9, pp. 636–645 (2016 Sep.), doi: 10.17743/jaes.2016.0032.

[29] L. Reymore, D. Huron, "Using auditory imagery tasks to map the cognitive linguistic dimensions of musical instrument timbre qualia," *Psychomusicology: Music, Mind, and Brain*, vol. 30, no. 3, pp. 124–144 (2020 Jun.), doi: 10.1037/pmu0000263.

[30] A. Zacharakis, J. Reiss, "An Additive Synthesis Technique for Independent Modification of the Auditory Perceptions of Brightness and Warmth," (2011 May).

[31] C. Saitis, K. Siedenburg, P. Schuladen, C. Reuter, "The role of attack transients in timbral brightness perception," presented at the *Proceedings of the 23rd International Congress on Acoustics*, p. 5506 (2019 Sep.).

[32] R. Ethington, B. Punch, "SeaWave: A system for musical timbre description," *Computer Music Journal*, vol. 18, no. 1, pp. 30–39 (1994 Jan.), doi:10.2307/3680520.

[33] A. Gounaropoulos, C. Johnson, "Synthesising Timbres and Timbre-Changes from Adjectives/Adverbs," presented at the *Applications of Evolutionary Computing. EvoWorkshops 2006*, pp. 664–675 (2006 Apr.), doi:10.1007/11732242_63.

[34] M. B. Cartwright, B. Pardo, "Social-EQ: Crowdsourcing an Equalization Descriptor Map." presented at the *Proceedings of the 14th International Society for Music Information Retrieval Conference*, pp. 395–400 (2013 Jan.).

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

[36] R. Stables, B. De Man, S. Enderby, J. Reiss, T. Wilmering, G. Fazekas, "Semantic description of timbral transformations in music production," presented at the *ACM Multimedia, Oct. 15-19, Amsterdam, Netherlands*, pp. 337–341 (2016 Oct.), doi:10.1145/2964284.2967238.

[37] P. Esling, N. Masuda, A. Bardet, R. Despres, A. Chemla-Romeu-Santos, "Flow Synthesizer: Universal Audio Synthesizer Control with Normalizing Flows," *Applied Sciences*, vol. 10, no. 1, p. 302 (2020 Dec.), doi: 10.3390/app10010302.

[38] J. M. Chowning, "The Synthesis of Complex Audio Spectra by Means of Frequency Modulation," *Journal of the Audio Engineering Society*, vol. 21, no. 7, pp. 526–534 (1973 Sep.), publisher: Audio Engineering Society.

[39] D. Wessel, "Control of phrasing and articulation in synthesis," presented at the *Proceedings of the 1987 International Computer Music Conference*, pp. 108–116 (1987).

[40] R. Ashley, "A knowledge-based approach to assistance in timbral design," presented at the *Proceedings of the 1986 International Computer Music Conference*, pp. 11–16 (1986 Oct.).

[41] Z. Wallmark, R. J. Frank, L. Nghiem, "Creating novel tones from adjectives: An exploratory study using FM synthesis." *Psychomusicology: Music, Mind, and Brain*, vol. 29, no. 4, pp. 188–199 (2019 Jul.), doi:10.1037/pmu0000240.

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

[43] J. Marozeau, A. de Cheveigné, "The effect of fundamental frequency on the brightness dimension of timbre," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 383–387 (2007 Jan.), doi:10.1121/1.2384910.

[44] D. Müllensiefen, B. Gingras, J. Musil, L. Stewart, "The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population," *PLOS ONE*, vol. 9, no. 2, p. e89642 (2014 Feb.), doi:10.1371/journal.pone.0089642, publisher: Public Library of Science.

[45] A. Zacharakis, B. Hayes, C. Saitis, K. Pastiadis, "Evidence for timbre space robustness to an uncontrolled online stimulus presentation," presented at the *Proceedings of the 2nd International Conference on Timbre* (2020 Sep.).

[46] F. Henninger, Y. Shevchenko, U. Mertens, P. J. Kieslich, B. E. Hilbig, "lab.js: A free, open, online experiment builder," (2020 Apr.), doi:10.5281/zenodo.3767907.

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

[48] H. F. Kaiser, "The application of electronic computers to factor analysis," *Educational and Psychological Measurement*, vol. 20, pp. 141–151 (1960 Apr.), doi: 10.1177/001316446002000116, place: US Publisher: Sage Publications.

[49] R. B. Cattell, "The Scree Test For The Number Of Factors," *Multivariate Behavioral Research*, vol. 1, no. 2, pp. 245–276 (1966), doi:10.1207/s15327906mbr0102_10.

[50] 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.), doi:10.1007/BF02289447.

[51] W. R. Zwick, W. F. Velicer, "Comparison of five rules for determining the number of components to retain," *Psychological Bulletin*, vol. 99, no. 3, pp. 432–442 (1986 May), doi:10.1037/0033-2909.99.3.432.

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

[53] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Tech. rep., IRCAM (2004 Jan.).

[54] J. Stepánek, "Musical sound timbre: Verbal description and dimensions," presented at the *Proceedings of the 9th International Conference on Digital Audio Effects (DAFx-06)*, pp. 121–126 (2006 Jan.).

[55] J. Noble, E. Thoret, M. Henry, S. Mcadams, "Semantic dimensions of sound mass music: mappings between perceptual and acoustic domains," *Music Perception: An Interdisciplinary Journal*, vol. 38, no. 2, pp. 214–242 (2020 Nov.), doi:10.1525/mp.2020.38.2.214.

[56] C. Saitis, 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.), doi:10.1121/10.0002275.

[57] M. Vowels, R. Mason, "Comparison of pairwise dissimilarity and projective mapping tasks with auditory stimuli," *Journal of the Audio Engineering Society*, vol. 68, no. 9, pp. 638–648 (2020 Sep.), doi:10.17743/jaes.2020.0051.

[58] S. S. Stevens, "Tonal Density," *Journal of Experimental Psychology*, vol. 17, no. 4, pp. 585–592 (1934).

[59] E. J. Allen, A. J. Oxenham, "Symmetric Interactions and Interference Between Pitch and Timbre," *The Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1371–1379 (2014 Mar.), doi:10.1121/1.4863269.

[60] A. B. Warriner, V. Kuperman, M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207 (2013 Feb.), doi:10.3758/s13428-012-0314-x.

[61] R. Vertegaal, E. Bonis, "ISEE: An Intuitive Sound Editing Environment," *Computer Music Journal*, vol. 18, no. 2, p. 21 (1994 Jan.), doi:10.2307/3680440.

[62] J. R. Lam, C. Saitis, "The Timbre Explorer: A Synthesizer Interface for Educational Purposes and Perceptual Studies," presented at the *NIME 2021* (2021 Jun.), doi:10.21428/92fbeb44.92a95683.

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

[64] M. Huzaifah, L. Wyse, "Deep generative models for musical audio synthesis," *arXiv:2006.06426 [cs, eess, stat]* (2020 Jun.), arXiv: 2006.06426.

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

[66] S. Huang, Q. Li, C. Anil, S. Oore, R. B. Grosse, "TimbreTron A WaveNet(CycleGAN(CQT(Audio))) Pipeline for Musical Timbre Transfer," presented at the *7th International Conference on Learning Representations*, p. 17 (2019 May).

[67] B. Hayes, C. Saitis, G. Fazekas, "Neural Waveshaping Synthesis," presented at the *Proceedings of the 22nd International Society for Music Information Retrieval Conference* (2021 Nov.).

[68] G. Bromham, D. Moffat, M. Barthet, A. Danielsen, G. Fazekas, "The Impact of Audio Effects Processing on the Perception of Brightness and Warmth," presented at the *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*, pp. 183–190 (2019 Sep.), doi: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 section 2.2.

	Description	Bigram Occ.	Corpus Occ.
1	great	12637	128040
2	good	6158	142535
3	nice	3584	92787
4	different	3271	80763
5	awesome	1896	32652
6	cool	1734	54245
7	amazing	1571	20479
8	interesting	1415	40124
9	fantastic	1286	9598
10	synth	1222	60582
11	percussive	1217	3482
12	pretty	1093	75287
13	similar	1089	29786
14	new	887	88297
15	unique	848	8253
16	beautiful	692	9237
17	digital	678	30144
18	clean	670	12526
19	complex	573	15652
20	incredible	555	4106
21	modular	552	118712
22	fm	540	27389
23	wonderful	536	6525
24	overall	516	5425
25	right	491	77903
26	bad	487	23048
27	weird	446	12432
28	excellent	446	11666
29	drum	437	41217
30	organic	419	2383
31	sweet	409	8992
32	crazy	408	11627
33	raw	385	3557
34	external	372	22864
35	natural	364	2684
36	fine	362	32489
37	basic	352	19560
38	classic	345	8470
39	original	330	23436
40	electronic	323	9695
41	much	322	120812
42	many	315	56465
43	huge	307	11131
44	rich	302	3003
45	big	300	34466
46	metallic	297	1268
47	musical	296	9838
48	specific	293	12207
49	decent	288	9692
50	certain	279	11501

A.2 Descriptor Pruning Criteria

1.Remove words referring to affect (e.g., *good*)
2.Remove words referring to specific synthesisers or hardware (e.g., *moogy*)

3.Keep only one element of any group of words sharing a stem, favouring 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 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, then gives us the overall loading matrix L with which we can specify models for X_c and X_r: $X_c = LF_c + M_c + \varepsilon_c$ and $X_r = LF_r + M_r + \varepsilon_r$. This loading matrix thus applies also to our model of observed comparative ratings:

$$\begin{aligned} X_{\text{diff}} &= X_c - X_r + \varepsilon_{\text{diff}} \\ &= L(F_c - F_r) + M_c - M_r + \varepsilon_c - \varepsilon_r + \varepsilon_{\text{diff}} \\ &= LF_{\text{diff}} + M_{\text{diff}} + \varepsilon. \end{aligned}$$

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.

A.4 Extracted Acoustic Features

Table 5 summarizes and briefly explains the extracted acoustic features.

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), UK, where he works under the supervision of Charalampos Saitis and György Fazekas as part of the UKRI Centre for Doctoral Training in Artificial Intelligence and Music. His research centres around novel applications of deep learning for modelling the synthesis and perception of musical timbre, with a particular focus on meta-learning techniques. He also holds an MSc degree in Sound and Music Computing from QMUL and a BMus(Hons) in Electronic Music from the Guildhall School of Music and Drama. He is an organising member of the Special Interest Group on Neural Audio Synthesis (SIGNAS) at C4DM, and in December 2021 he organised the first international Neural Audio Synthesis Hackathon (NASH). 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 PhD 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 crossmodality, and "metaphors we listen with". He acted as co-editor of the Springer Series on Touch and Haptic Systems volume on Musical Haptics (2018) and the Springer Handbook of Auditory Research volume on Timbre (2019), and has authored several recent publications on timbre perception and semantics. He was co-organiser of the Berlin Interdisciplinary Workshop on Timbre (2017) and a founding member of the International Conference on Timbre (2020).

Table 5. Extracted acoustic features

Signal Representation	*Feature*	*Explanation*
$STFT_{mag}$ Spectrum	Centroid	Centre 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 centre 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

●

George Fazekas is a Senior Lecturer at the Center for Digital Music, Queen Mary University of London. He holds a BSc, MSc and PhD degree in Electronic Engineering. He is an investigator of UKRI's £6.5M Centre for Doctoral Training in Artificial Intelligence and Music (AIM CDT) and he was QMUL's Principal Investigator on the H2020 funded Audio Commons project. He was general chair of ACM'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.