There's more to timbre than musical instruments: semantic dimensions of FM sounds

Ben Hayes[1†], Charalampos Saitis[1]
[1] Centre for Digital Music, Queen Mary University of London, London, United Kingdom
[†] Corresponding author: b.j.hayes@se19.qmul.ac.uk

Introduction

Much previous research into timbre semantics (such as when an oboe is described as "hollow") has focused on sounds produced by acoustic instruments, particularly those associated with western tonal music (Saitis & Weinzierl, 2019). Many synthesisers are capable of producing sounds outside the timbral range of physical instruments, but which are still discriminable by their timbre. Research into the perception of such sounds, therefore, may help elucidate further the mechanisms underpinning our experience of timbre in the broader sense. In most timbre semantics research, listeners rate a set of sounds along scales defined by descriptive adjectives. By reverse engineering the standard paradigm, a smaller number of studies have provided evidence that musicians can manipulate timbre in abstract synthesis scenarios to match certain adjective descriptions. For example, Wallmark *et al.* (2019) employed a simplified FM (Frequency Modulation) synthesis interface to study the relationship between semantic descriptors and sound creation, showing an association between word valence and specific acoustic features.

In this paper, we present a novel paradigm on the application of semantic descriptors to sounds produced by experienced sound designers using an FM synthesiser with a full set of controls. FM synthesis generates rich and complex timbres via time-varying phase modulation of sinusoidal oscillators (Chowning, 1973), and is amenable to statistical analysis as broad timbral palettes can be expressed as a function of a completely continuous parameter space. Our aim with this work is twofold. First, we intend to ascertain whether the luminance-texture-mass (LTM) model of timbre semantics (Zacharakis *et al.*, 2014) is sufficient to describe the semantic dimensions of sounds produced through FM synthesis. Secondly, we hope the collected data and subsequent analysis will form a basis for future work into perceptually informed deep-learning based semantic synthesis control schemes.

Method

Thirty participants[1] completed the experiment (mean age 28.7 years; range 21-55 years). All spent their formative years in an English speaking country and self-reported having prior experience with synthesis through either music production or sound design backgrounds. Owing to the infeasibility of conducting an in-person study during the COVID-19 pandemic, the study took place online using the WebAudio API to generate sounds and the *lab.js* framework to collect data (Henniger *et al.*, 2020)[2].

The synthesiser employed a three operator architecture, with operators 2 and 3 modulating the phase of operator 1 in linear combination. Each operator had an independent *attack-decay-sustain-release* envelope. Participants were presented with a browser-based synthesiser interface with controls pre-set to generate a reference sound. An instruction was given to adjust the parameters such that the synthesiser produced a new sound matching a given comparative prompt (e.g. 'brighter' or 'less thick'). Each participant undertook nine trials covering each combination of three LTM prompts (*bright*, *rough, thick*) and three pitches (E2, A3, D5). Each trial, the positive or negative comparative form of the relevant prompt was selected randomly. Participants were then asked to rate the magnitude of the difference between the two sounds in terms of the prompt, as well as the difference between the created sound and the reference sound in terms of the remaining two LTM descriptors and an additional set of 24 semantic descriptors.

[1] Forty took part in total, but 10 were not used for analysis due to not meeting age or language restrictions
[2] Source code for the study is available in a GitHub repository: https://github.com/ben-hayes/fm-synth-study

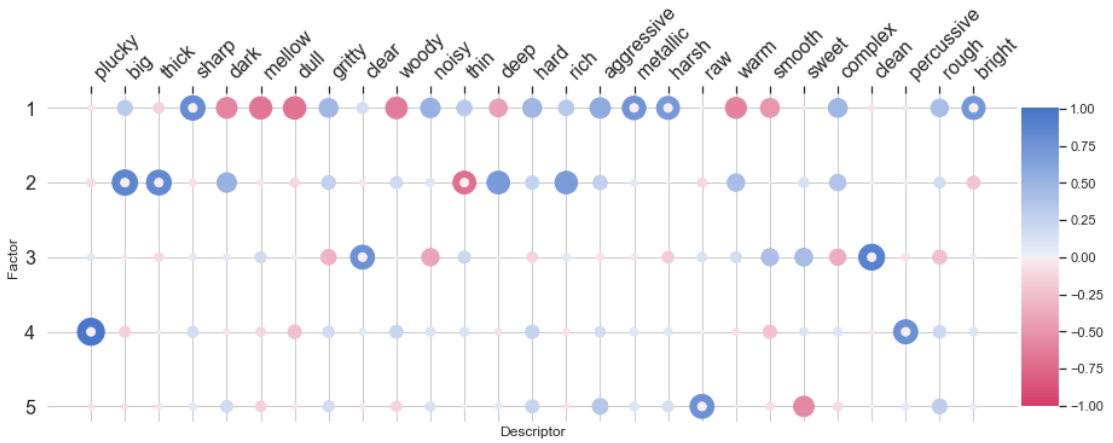

Figure 1: The factor loadings of the five factor solution across all descriptors. A white dot in the centre of a point indicates an absolute loading ≥ 0.7

Descriptors were selected for the experiment by mining and analysing a text corpus from the popular synthesis forum *MuffWiggler*. This approach was selected to maximise appropriateness to the sonic domain of synthesised sounds. The corpus was filtered to a frequency-sorted list of words co-occurring in bigrams with the terms 'sound', 'sounding', 'tone', and 'timbre', and this list was filtered so that only the top 100 adjectives remained. These were independently pruned by two raters according to a set of criteria, resulting in the final set of 27 descriptors. The LTM prompts were selected as the descriptors with the highest corpus frequencies that also showed significant loadings onto the English LTM factors in Zacharakis *et al.* (2014). To avoid biasing the semantic responses towards the characteristics of a fixed set of starting sounds, it was deemed advantageous to explore participants' responses across as much of the synthesiser's parameter space as possible. To this end, the reference sounds presented in each trial were randomly selected from the database of sounds created by previous participants, with the proviso that sounds may not traverse pitch conditions. As well as presenting a more balanced representation of the sonic properties of the synthesis method and participants' responses to LTM prompts, this approach confers the additional benefit that the resulting dataset is more amenable to future use in deep learning synthesis models.

An exploratory factor analysis with non-orthogonal oblimin rotation was performed on the resulting comparative descriptor ratings, using the maximum-likelihood method. The number of factors was selected using parallel analysis (Horn, 1965), a procedure in which the eigenvalues of the data correlation matrix are compared to those of a large number of correlation matrices generated from normally distributed random datasets via a Monte-Carlo simulation. The number of factors then corresponds to the number of eigenvalues from the real data's correlation matrix that exceed a given percentile (typically the 95th) of the synthetic data's eigenvalues. Parallel analysis has been shown to outperform the Kaiser method of retaining factors with eigenvalues greater than 1.0 (Zwick & Velicer, 1986). Finally, the monotonicity of the relationships between synthesiser parameter changes and descriptive prompts were studied by computing the Spearman rank correlation.

Results

Factor Analysis

Performed on all descriptors across all prompts, parallel analysis supported a five factor solution, using the 95th percentile as a threshold. The resulting factors cumulatively accounted for 74.36% of data variance. Fig. 1 illustrates descriptor loadings onto the rotated factors. Notably, factor 1 shows strong loadings onto terms associated with luminance (including *sharp*) as well as terms associated with texture (*metallic, harsh)*. Factor 2 shows strong loadings onto terms related to mass (*big*, *thick*, and negatively *thin*). Factor

Figure 2: Correlations between semantic factors.

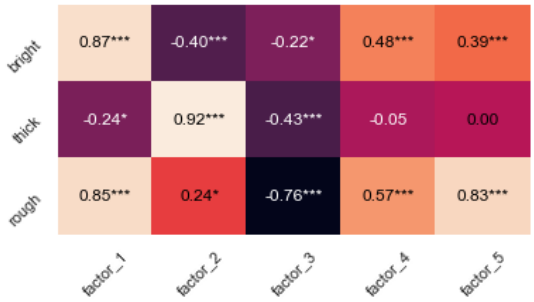

Figure 3: Pearson correlation coefficients between prompts and semantic factors. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

3 shows strong loadings for words associated with clarity (*clean, clear*), factor 4 for "pluckiness" (*plucky, percussive*), and factor 5 for *raw*. Fig. 2 shows the inter-factor correlations after non-orthogonal Oblimin rotation. Here we see moderate collinearity between factors, most notably between factor 1 and factors 3-5. Fig. 3 shows the correlations between reported prompt magnitudes and semantic factors within each prompt condition. Each row, therefore, represents a non-overlapping subset of the dataset as each created sound was prompted by only one of the three LTM prompts. Factor 1 shows strong and significant correlations with *bright* and *rough* prompt magnitudes, and factor 2 with *thick.* Factors 3-5, however, all exhibit markedly different relationships with the prompts.

Parameter Correlations

Fig. 4 illustrates the Spearman rank correlation coefficients between reported prompt magnitudes and changes to synthesiser parameters within each prompt condition. The *bright* and *rough* prompts express very similar patterns of correlations, which both imply a tendency to, in response a positive prompt, increase the gains and tuning ratios of modulating operators (thereby increasing the energy and frequency of sidebands, respectively), and to decrease the attack time of the carrier operator (which decreases the overall attack time of the sound; cf. Saitis *et al.*, 2019). The most significant correlations observed for the *thick* prompt occur with the three sustain parameters, with the strongest of which was with the carrier operator sustain (which decreases the attenuation of the sustain portion of the sound).

Discussion

Comparing the loadings (Fig. 1) of factor 1 to those found by Zacharakis *et al.*, (2014) for terms present in both studies suggests it is an amalgamation of luminance and texture dimensions. The patterns of parameter delta correlations shared by the *bright* and *rough* prompts (Fig. 2) suggest that this is a direct result of the properties of FM synthesis: it is challenging to increase the energy in high frequency components (by increasing the modulator tuning or gain) without also increasing inharmonicity. The

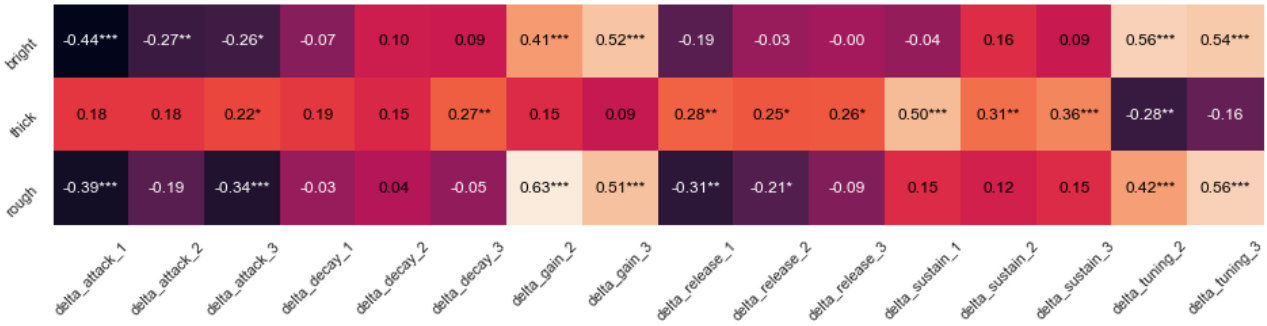

Figure 4: The Spearman rank correlation coefficients between prompt magnitudes and changes to synthesiser parameters. $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).

loadings (Fig. 1) of factor 2 resemble the mass dimension of Zacharakis *et al.*, (2014). Factors 1-2 therefore appear to reflect the LTM prompts used for sound design and, indeed, the strong and significant correlations between these factors and the corresponding prompts (Fig. 3) support this hypothesis.

The loadings (Fig. 1) of factors 3-5 (*clean*, *plucky*, *raw*) suggest the created sounds exhibit attributes not entirely accounted for by LTM factors, and the correlations (Fig 3.) between these factors and prompt magnitudes support this. However, the moderate correlations between these factors and factors 1-2 (Fig. 2) suggest that these attributes are not entirely independent. Ascertaining whether this is due to an inherent property of the synthesiser or the interpretation of the descriptors themselves is left to future analysis.

Conclusions & Future Work

In this study we presented a novel paradigm for studying both the response of experienced sound designers to semantic prompts, and the semantic dimensions of the sounds they created. Exploratory factor analysis yielded a five factor model of which the first two factors correspond to the factors of the LTM model (factor 1: joint luminance-texture, factor 2: mass). The extra three factors appear to correspond, respectively, to clarity, pluckiness, and rawness. In subsequent analysis, acoustic features will be extracted from all synthesiser patches created in the study, enabling the psychoacoustic underpinnings of the semantic space to be analysed and, in particular, the relationship between factors 3-5 and factors 1-2. Owing to the design of this experiment — in particular, the fact that each stimulus is rated by only a single participant, and the use of comparative semantic ratings — it will be necessary to confirm its efficacy and the structure of the resulting semantic space with a classical semantic rating design (Zacharakis *et al.,* 2014) in order to compare the resulting factors to those found in previous studies.

Research into the semantics and perception of synthesised sounds provides a basis for future work into enhanced approaches for the control of synthesisers. Integration with semantic audio technologies and application of neural audio synthesis techniques will enable the intuitive generation and manipulation of novel timbres. Further, continued study of the broad and abstract sonic palettes afforded by synthesis methods such as FM will enable deeper insight into the intrinsic properties of timbre, as opposed to only those associated with physical sources, allowing for a more complete conception of the mechanisms underpinning its perception.

References

Chowning, J. M. (1973). The Synthesis of Complex Audio Spectra by Means of Frequency Modulation. *Journal of the Audio Engineering Society*, *21*(7), 526–534.

Henninger, F., Shevchenko, Y., Mertens, U., Kieslich, P. J., & Hilbig, B. E. (2020). *lab.js: A free, open, online experiment builder*. Zenodo.

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

Saitis, C., Siedenburg, K., Schuladen, P., and Reuter, C. (2019). The role of attack transients in timbral brightness perception. In: Vorländer M., Fels J. (eds), *Proceedings of the 23rd International Congress on Acoustics*, (pp. 5506-5543). Aachen, Germany.

Saitis, C., & Weinzierl, S. (2019). The semantics of timbre. In K. Siedenburg, C. Saitis, S. McAdams, *et al.* (eds.), *Timbre: Acoustics, Perception, and Cognition* (pp. 119–149). Springer Handbook of Auditory Research, vol 69. Springer, Cham.

Wallmark, Z., Frank, R. J., & Nghiem, L. (2019). Creating novel tones from adjectives: An exploratory study using FM synthesis. *Psychomusicology*, 29 (4), 188–199.

Zacharakis, A., Pastiadis, K., & Reiss, J. D. (2014). An Interlanguage Study of Musical Timbre Semantic Dimensions and Their Acoustic Correlates. *Music Percept., 31 (4)*, 339–358.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99* (3), 432–442.