Semantic Description of Timbral Transformations in Music Production

Ryan Stables
Digital Media Technology Lab
Birmingham City University
Birmingham, UK
ryan.stables@bcu.ac.uk

Brecht De Man
Centre for Digital Music
Queen Mary University of London
London, UK
b.deman@qmul.ac.uk

Sean Enderby
Digital Media Technology Lab
Birmingham City University
Birmingham, UK
sean.enderby@mail.bcu.ac.uk

Joshua D. Reiss
Centre for Digital Music
Queen Mary University of London
London, UK
joshua.reiss@qmul.ac.uk

György Fazekas
Centre for Digital Music
Queen Mary University of London
London, UK
g.fazekas@qmul.ac.uk

Thomas Wilmering
Centre for Digital Music
Queen Mary University of London
London, UK
t.wilmering@qmul.ac.uk

ABSTRACT

In music production, descriptive terminology is used to define perceived sound transformations. By understanding the underlying statistical features associated with these descriptions, we can aid the retrieval of contextually relevant processing parameters using natural language, and create intelligent systems capable of assisting in audio engineering. In this study, we present an analysis of a dataset containing descriptive terms gathered using a series of processing modules, embedded within a Digital Audio Workstation. By applying hierarchical clustering to the audio feature space, we show that similarity in term representations exists within and between transformation classes. Furthermore, the organisation of terms in low-dimensional timbre space can be explained using perceptual concepts such as size and dissonance. We conclude by performing Latent Semantic Indexing to show that similar groupings exist based on term frequency.

CCS Concepts

•**Information systems** → **Information systems applications; Multimedia information systems;** *Multimedia databases;*

Keywords

Semantic Audio, Timbre, Music Production, Hierarchical Clustering, Dimensionality Reduction

1. INTRODUCTION

Musical timbre refers to the properties of a sound, other than loudness and pitch, which allow it to be distinguished from other sounds [8]. Loudness and pitch can easily be measured in low-dimensional space, allowing sounds to be ordered from quiet to loud or low to high in frequency, whereas timbre is a more complex property of sound, requiring multiple dimensions [11]. To characterise perceptual attributes of musical timbre, listeners often attribute semantic descriptors such as *bright*, *rough* or *sharp* to describe latent dimensions [5]. A widely cited definition of timbre [1] shows it can be determined by a range of low level features of an audio signal, where the spectral content and temporal characteristics both affect the perceived timbre of a sound. Signal analysis techniques can be used to extract information about these elements of a signal. The contribution of these low level features to perceived timbre is often the focus of academic research, whereby dimensionality reduction techniques allow for the organisation of terms in an underlying subspace, with the intention of discovering some perceptually relevant representation of the data [2, 4, 6, 17]. In music production, this is of particular interest as it can allow for the manipulation of audio processing modules, comprising multiple parameters using intuitive, low-dimensional controls [3, 12, 14, 15].

In this paper we report our findings from the Semantic Audio Feature Extraction (SAFE) Project [13], and show that semantic descriptions of musical timbre can be grouped using both parameter and feature space representations, and can exhibit timbral similarities within and across audio processing types. We investigate the use of timbral descriptors to aid the retrieval of contextually relevant processing parameters given natural language descriptions of audio transformations. This allows for the development of intuitive and assistive music production interfaces, based on descriptive cues.

2. SAFE

The Semantic Audio Feature Extraction (SAFE) plug-ins[1] provide music producers with a platform to describe timbral transformations in a Digital Audio Workstation (DAW) using natural language [13]. The plugins (referred to herein as transform classes) consist of a five band parametric equaliser,

[1]Plugins and datasets available at semanticaudio.co.uk.

MM '16, October 15 - 19, 2016, Amsterdam, Netherlands

ISBN 978-1-4503-3603-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2964284.2967238

	Num Instances		Confidence		Popularity		Generality	
N	term	n	term	c	term	p	term	g
0	warm	193	boxed	.250	warm	0.0019	sharp	.828
1	bright	153	splash	.250	bright	0.0014	deep	.819
2	punch	34	wholesome	.250	crunch	0.0006	boom	.809
3	air	31	pumping	.247	room	0.0005	thick	.806
4	crunch	29	rounded	.247	fuzz	0.0004	piano	.696
5	room	28	sparkle	.247	crisp	0.0004	strong	.596
6	smooth	22	atmosphere	.244	clear	0.0004	soft	.575
7	vocal	21	balanced	.244	cut	0.0004	bass	.555
8	clear	20	bass	.244	bass	0.0004	gentle	.525
9	fuzz	19	basic	.244	low	0.0004	tin	.483

Table 1: The highest ranking terms using confidence, popularity and generality measures.

a dynamic range compressor, amplitude distortion and a reverb effect. When a timbral transformation is recorded, the system extracts the descriptive terminology relating to the transform; a large set of temporal, spectral and abstracted audio features taken across a number of frames of the audio signal, both before and after processing (see [9] for a full list); the name and parameter settings of the audio effect; and a list of additional user data such as age, location, production experience, genre and instrument. This information is stored in an RDF triple store using an empirically designed ontology.

2.1 Dataset

The dataset used for the study comprised 2694 transforms, split into four groups according to their transform class. Overall, 454 were applied using a compressor, 303 using distortion, 1679 using an equaliser, and 258 using a reverb. The transforms were described using 618 unique terms taken from 263 unique users (averaging 2.35 terms per user), all of whom were music producers who participated by using the SAFE Plugins within their workflow.

We measure the *confidence* of a descriptor using the sum of its variance in feature space, where each of the features is mapped to a 6-dimensional space using Principal Component Analysis (PCA) in order to remove redundancy, whilst retaining $\geq 95\%$ of the variance:

$$c = \frac{1}{M} \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} (PC_n(m) - \mu_n)^2 \quad (1)$$

To further identify the *popularity* of a descriptor, we weight the output of Eq. (1) with a coefficient representing the term as a proportion of the dataset:

$$p = c \cdot ln \frac{n(d)}{\sum_{d=0}^{D-1} n(d)} \quad (2)$$

where $n(d)$ is the number of entries for a descriptor d.

Finally, we evaluate the extent to which the descriptor is generalisable across a range of transform classes (*generality*) by finding the weighted mean of the term's sorted distribution. This is equivalent to finding the centroid of the density function across transform classes.

$$g = \frac{2}{K-1} \sum_{k=0}^{K-1} k \, \mathrm{sort}(x(d))_k \quad (3)$$

where the distribution of the term d is calculated as a proportion of the transform class (k) to which it belongs:

$$x(d)_k = \frac{n_d(k)}{N(k)} \frac{1}{\sum_{k=0}^{K-1} N(k)} \quad (4)$$

Here, $N(k)$ is the total number of entries in class k and $n_d(k)$ is the number of occurrences of descriptor d in class k. Using these metrics, the database is sorted and the top 10 descriptors are shown in Table 1. Similarly, Table 2 shows the most commonly used descriptors for each individual transform class. To group terms with shared meanings and variable suffixes, stemming conditions are applied using a Porter Stemmer [10]. This allows for the unification of terms such as *warm, warmer and warmth* into a parent category (*warm*).

Compressor	Distortion	EQ	Reverb
27 : punch	23 : crunch	440 : warm	30 : room
17 : smooth	20 : warm	424 : bright	13 : air
15 : sofa	6 : fuzz	16 : air	11 : big
14 : vocal	6 : destroyed	16 : clear	10 : subtle
12 : nice	5 : cream	12 : thin	9 : hall
9 : controlled	5 : death	11 : clean	9 : small
9 : together	5 : bass	11 : crisp	8 : dream
9 : crushed	5 : clip	10 : bass	7 : damp
8 : warm	5 : decimated	9 : boom	7 : drum
7 : comp	5 : distorted	9 : cut	6 : close

Table 2: The first ten descriptors per class, ranked by number of entries.

3. WITHIN-CLASS SIMILARITY

To find term-similarities within transform classes, hierarchical clustering is applied to differences (processed vs. unprocessed) in timbre space. To do this, the mean of the audio feature vectors from each unique descriptor is computed and PCA is applied, reducing the number of dimensions, whilst preserving $\geq 95\%$ of the variance. Terms with < 8 entries are omitted for readability and the distances between datapoints are calculated using Ward distance [16], the results of which are shown in Figure 1. In each transform class, clusters are intended to retain perceived latent groupings, based on underlying semantic representations.

From the term clusters, distances between groups of semantically similar timbral descriptions emerge. Among the Compressor terms, groups tend to exhibit correlation with the extent to which gain reduction is applied to the signal. *Loud*, *fat* and *squashed* generally refer to extreme compression, whereas *subtle*, *gentle* and *soft* tend to describe minor adjustments to the amplitude envelope. Distortion features tend to group based on the perceived dissonance of the

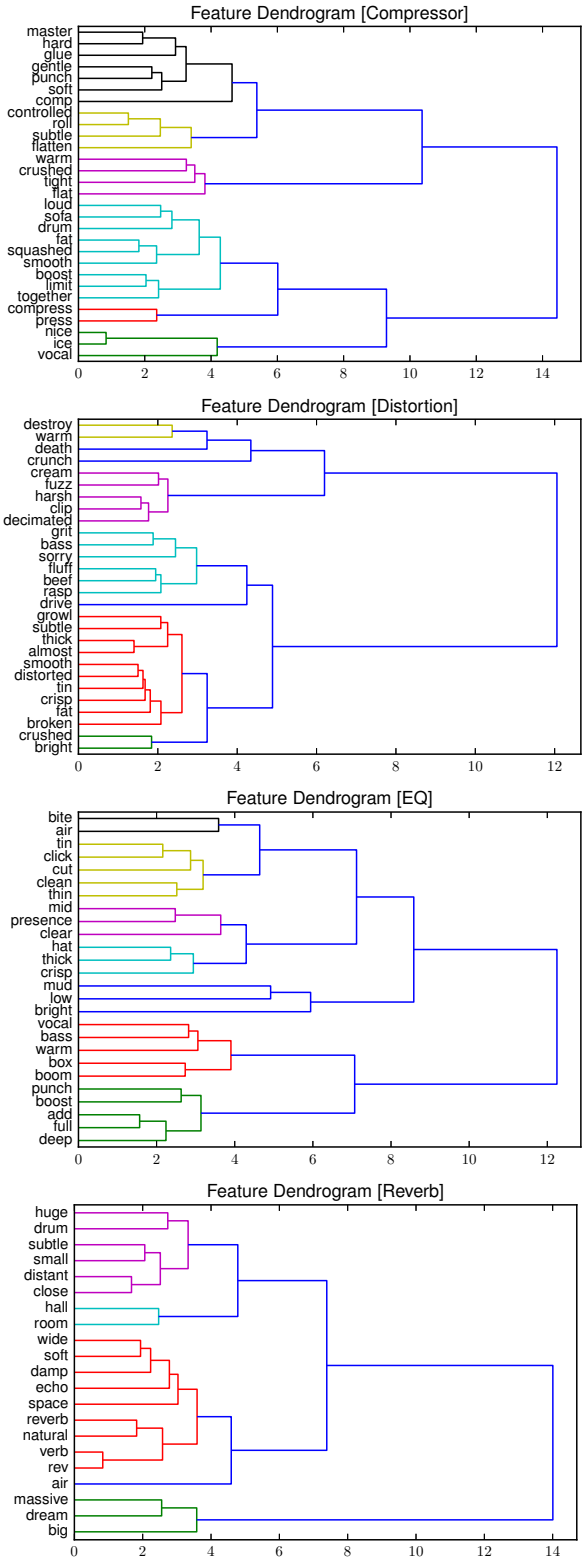


Figure 1: Dendrograms showing clustering based on feature space distances for each transform class.

transform, with terms such as *fuzz* and *harsh* clearly separated from *subtle*, *rasp* and *growl*. Equalisation comprises a wide selection of description-categories, although terms that generally refer to specific regions of spectral energy such as *bass*, *mid* and *full* tend to fall into separate partitions. Reverb terms tend to group based on size and descriptions of acoustic spaces. *Hall* and *room* for example exhibit similar feature spaces, while terms such as *soft*, *damp* and *natural* fall into the same group.

3.1 Parameter Space Representation

To illustrate the relevance of the within-class feature groups found using the hierarchical clustering algorithm, we can show that terms within clusters maintain similar characteristics in their parameter spaces. To demonstrate this, Figure 2 shows curves corresponding to two groups of descriptors taken from opposing clusters in the equaliser's feature-space: cluster 2 (*warm*, *bass*, *boom*, *box* and *vocal*) and cluster 8 (*thin*, *clean*, *cut*, *click* and *tin*). Curves in cluster 2 generally exhibit a boost around 500 Hz with a high-frequency roll-off, whereas terms in cluster 8 exhibit a boost in high-frequency energy centered around 5 kHz.

To further evaluate the organisation of terms based on their position in a parameter space, we use PCA to reduce the dimensionality of each space and overlay the parameter vectors. Figure 3 shows this for the distortion and reverb, where in 3(a) the bias is highly correlated with PC2, which tends to organise descriptors based on dissonance. Similarly in 3(b), the mix and gain parameters of the reverb class correlate with PC2 and tend to retain variance using size-based descriptors. These exhibit 0.68 and 0.81 cross-correlation values respectively.

4. INTER-TRANSFORM SIMILARITY

To investigate between-class similarities, we perform hierarchical clustering on the dataset, where transforms are grouped by unique terms and separated by transform class. Here, the organisation of terms into clusters is highly correlated with the organisation of terms into transform classes. Out of the 8 data partitions, the mean rank-order generality is 0.23, with a mean of 2.4 unique class labels per group.

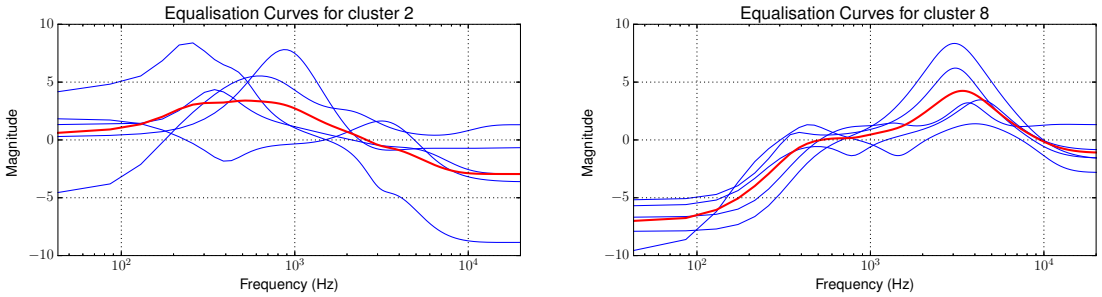
To identify transform-agnostic descriptors, i.e. those with similar between-class transformations, we select the top 10 terms with the highest generality scores (defined in Table 1) and measure the variance across the transformations in reduced-dimensionality space. All terms had entries in all 4 transform classes, and had at least 10 entries overall. Ranked by between-class agreement: 1. *piano* (0.001), 2. *sharp* (0.012), 3. *soft* (0.013), 4. *thick* (0.018), 5. *tin* (0.021), 6. *deep* (0.022), 7. *bass* (0.033), 8. *gentle* (0.039), 9. *strong* (0.050), 10. *boom* (0.058).

4.1 Term Frequency Analysis

We measure term similarity independently of timbral or parameter space representations, using a term's association to a given transform class. Here, we use term frequency to define distributions across classes, resulting in four-dimensional vectors, e.g. $\mathbf{t} = [0.0, 0.5, 0.5, 0.0]$ has equal association with the distortion and equaliser, but no entries in the compressor or reverb classes. We then represent these using a Vector Space Model (VSM), and measure similarity between any two terms $(\mathbf{t_1}, \mathbf{t_2})$ using cosine distance:

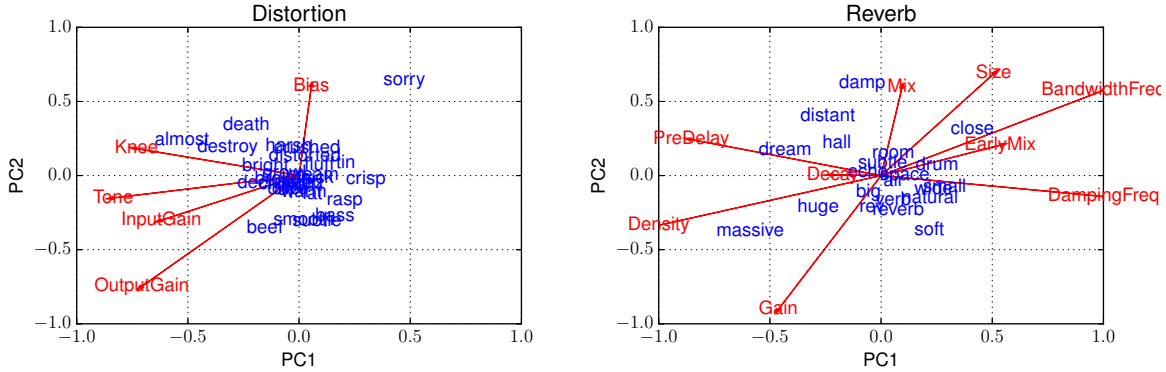
$$\text{sim}(\mathbf{t_1}, \mathbf{t_2}) = \frac{\mathbf{t_1} \cdot \mathbf{t_2}}{\|\mathbf{t_1}\|\|\mathbf{t_2}\|} = \frac{\sum_{i=1}^{N} t_{1,i} t_{2,i}}{\sqrt{\sum_{i=1}^{N} t_{1,i}^2}\sqrt{\sum_{i=1}^{N} t_{2,i}^2}} \quad (5)$$

In order to better capture the true semantic relations of



(a) *warm*, *bass*, *boom*, *box* and *vocal* (b) *thin*, *clean*, *cut*, *click* and *tin*

Figure 2: Equalisation curves for two clusters of terms in the dataset.



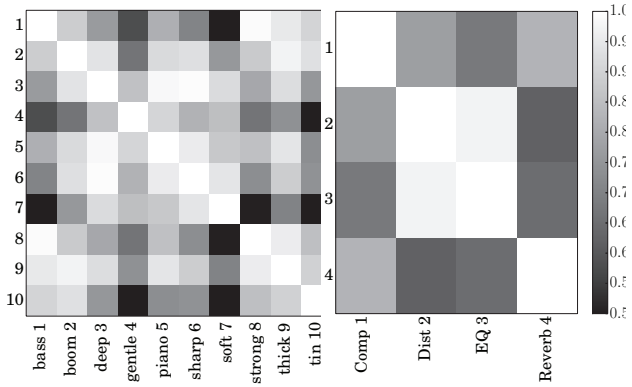
(a) (b)

Figure 3: Biplots of the distortion and reverb classes, showing terms mapped onto 2 dimensions with overlaid parameter vectors.

the terms and the transforms they are associated with, we apply Latent Semantic Indexing (LSI) [7], a process that involves reducing the term-transform space from rank four to three by performing a singular value decomposition of the $N_{terms} \times 4$ occurrence matrix $\mathbf{M} = \mathbf{U\Sigma V}^*$, and setting the smallest singular values to zero before reconstructing it using $\mathbf{M}' = \mathbf{U\Sigma}'\mathbf{V}^*$.

This process eliminates noise caused by differences in word usage, for instance due to synonymy and polysemy, whereas the 'latent' semantic relationships between terms and effects are preserved. Figure 4 shows the resulting pairwise similarities of the high-generality terms used in Section 4.

Here, the most similar terms are *bass* and *strong*, *deep* and *sharp* and *boom* and *thick* (all 0.99). Conversely, we can consider the similarity of transform types based on their descriptive attributes by transposing the occurrence matrix in the VSM. This is illustrated in Figure 4, in which terms used to describe equalisation transforms are similar to those associated with distortion (0.95), while equalisation and compression vocabulary is disjunct (0.641).



(a) (b)

Figure 4: Vector-space similarity wrt. (a) high-generality terms and (b) transform-classes.

5. DISCUSSION/CONCLUSION

We have illustrated within- and between-class groupings of semantic descriptions of sound transformations taken from processing modules in a DAW. We showed that the groups represent meaningful subsets of entries by evaluating correlation in their parameter spaces, and that the parameters of each processing module can be used to organise terms in a similar fashion. To evaluate between-transform similarity, we demonstrated that transforms tend to form the basis of discrete clusters, and that terms such as *piano*, *sharp*, *soft*, *thick* and *tinny* have similar representations across a range of processing types. Finally, we measured the similarity of effects and terms based on their vector-space representations. This shows that equalisation and compression share a common vocabulary of terms, whilst reverb and distortion have a dissimilar description schema. The results are encouraging and show that timbre descriptors cluster in meaningful ways in the context of audio transformations. The findings thus provide useful insight into how to create semantic descriptor spaces for audio effects.

6. REFERENCES

[1] American Standards Association. American standard acoustical terminology (including mechanical shock and vibration). Technical report, 1960.

[2] A. Caclin, S. McAdams, B. Smith, and S. Winsberg. Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *The Journal of the Acoustical Society of America*, 118(1):471–482, 2005.

[3] M. B. Cartwright and B. Pardo. Social-EQ: Crowdsourcing an equalization descriptor map. In *14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013.

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

[5] D. Howard and J. Angus. *Acoustics and Psychoacoustics*. Focal Press, 4th edition, 2009.

[6] R. Kendall and E. Carterette. Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives. *Music Perception: An Interdisciplinary Journal*, 10(4):445–467, 1993.

[7] T. A. Letsche and M. W. Berry. Large-scale information retrieval with latent semantic indexing. *Information sciences*, 100(1):105–137, 1997.

[8] M. Mathews. Introduction to timbre. In P. Cook, editor, *Music, Cognition, and Computerized Sound: An Introduction to Psychoacoustics*, chapter 7. MIT Press, 1999.

[9] G. Peeters. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical report, IRCAM, 2004.

[10] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[11] T. Rossing, R. Moore, and P. Wheeler. *The Science of Sound*. Addison Wesley, 3 edition, 2002.

[12] P. Seetharaman and B. Pardo. Socialreverb: crowdsourcing a reverberation descriptor map. In *ACM International Conference on Multimedia*, November 2014.

[13] R. Stables, S. Enderby, B. De Man, G. Fazekas, and J. Reiss. SAFE: A system for the extraction and retrieval of semantic audio descriptors. In *15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[14] S. Stasis, R. Stables, and J. Hockman. A model for adaptive reduced-dimensionality equalisation. In *18th International International Conference on Digital Audio Effects (DAFx-15), Trondheim, Norway*, 2015.

[15] S. Stasis, R. Stables, and J. Hockman. Semantically controlled adaptive equalisation in reduced dimensionality parameter space. *Applied Sciences*, 6(4):116, 2016.

[16] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[17] A. Zacharakis, K. Pastiadis, J. D. Reiss, and G. Papadelis. Analysis of musical timbre semantics through metric and non-metric data reduction techniques. In *12th International Conference on Music Perception and Cognition (ICMPC)*, pages 1177–1182, 2012.