OPEN ACCESS Freely available online
D. Moffat, B. De Man, and J. D. Reiss, "Semantic Music Production: A Meta-Study"
J. Audio Eng. Soc., vol. 70, no. 7/8, pp. 548–564, (2022 July/August).
DOI: https://doi.org/10.17743/jaes.2022.0023

Semantic Music Production: A Meta-Study

DAVID MOFFAT,[1,*] *AES Member*, **BRECHT DE MAN,**[2] *AES Member* **AND**
(dmof@pml.ac.uk) (brecht.deman@pxl.be)
JOSHUA D. REISS,[3] *AES Fellow*
(joshua.reiss@qmul.ac.uk)

[1]*Plymouth Marine Laboratory, Plymouth, UK*
[2]*PXL-Music, PXL University of Applied Sciences and Arts, Hasselt, Belgium*
[3]*Centre for Digital Music, Queen Mary University of London, London, UK*

This paper presents a systematic review of semantic music production, including a meta-analysis of three studies into how individuals use words to describe audio effects within music production. Each study followed different methodologies and stimuli. The SAFE project created audio effect plug-ins that allowed users to report suitable words to describe the perceived result. SocialFX crowdsourced a large data set of how non-professionals described the change that resulted from an effect applied to an audio sample. The Mix Evaluation Data Set performed a series of controlled studies in which students used natural language to comment extensively on the content of different mixes of the same groups of songs. The data sets provided 40,411 audio examples and 7,221 unique word descriptors from 1,646 participants. Analysis showed strong correlations between various audio features, effect parameter settings, and semantic descriptors. Meta-analysis not only revealed consistent use of descriptors among the data sets but also showed key differences that likely resulted from the different participant groups and tasks. To the authors' knowledge, this represents the first meta-study and the largest-ever analysis of music production semantics.

0 INTRODUCTION

Semantic analysis enables novel methods of interfacing with music production technologies, in which users can manipulate the processing by simply describing perceptual changes to be applied, for example, "make this sound 'brighter.'" Three studies attempted to gather a large number of terms from a large number of users, such that in-depth analysis could be performed [1–3]. Together, the data sets provided a total of 40,411 audio examples of 7,221 unique descriptors from 1,646 participants. Each of these studies used different methodologies, stimuli, and participants. This presents an opportunity for cross-analysis to determine whether the semantics is maintained with different groups of people, for different content and different music production tasks. It also allows aggregation to deliver the first meta-study of music production semantics and the largest study of music semantics to date.

The SocialFX study, presented in [1], is a combined study in which participants were asked to listen to an audio track both before and after an audio effect was applied. They were asked to both free-text words that describe the effect and select words from a short list, which had previously been submitted by other participants, as a form of validation or agreement. The study was conducted as an online listening test, through Amazon Mechanical Turk, where participants were paid to take part in the study. A total of 481 participants were included for equalization, 513 participants for reverberation, and 239 participants for dynamic range compression. It is not known whether there was any intersection between participants in each of these studies.

The SAFE data set, presented in [2], developed a range of four open-source audio effect Virtual Studio Technology plugins, equalizer, reverb, dynamic range compressor, and distortion effect. Participants were then provided the plugin, and they could change any effect parameters on any audio

*Correspondence should be addressed to David Moffat, e-mail: dmof@pml.ac.uk

track and were free to describe the audio effect using any words that they felt suitable. Participants were also able to recall effect parameters from the word-parameter database. There were 263 unique users of the audio effects.

The Mix Evaluation Data Set (MEDS), presented in [3], collected 181 different mixes of 18 different songs, produced by 150 unique individuals. Subjective evaluations of the mixes were then produced multiple times for each mix, which included a free-text description of the mix and a rating, when compared with other mixes of the same song. The free-text annotations were then analyzed and summarized in [3], to relevant individual semantic terms relating to the entire mix. All data sets presented are openly available. A more in-depth review of each of the three data sets is presented in SEC. 2

Herein, this paper presents a meta-analysis of these studies. Analysis showed strong correlations between various audio features, effect parameter settings, and semantic descriptors. Meta-analysis not only revealed consistent use of descriptors among the data sets but also showed key differences that likely resulted from the different participant groups and variations within the completed tasks.

The paper is organized as follows. SEC. 1 reviews the field of semantic music production. SEC. 2 describes in further detail the three data sets that were used. SEC. 3 presents a secondary analysis based on combination of the three data sets. SEC. 4 describes the meta-analysis approach that was taken and presents the results of this analysis. SEC. 5 discusses the findings, including a critique of the work. Finally, SEC. 6 presents overall conclusions, recommendations, and directions for further work.

1 BACKGROUND

Semantics can be introduced or applied in music production in the following ways:

- Creating music with meaningful controls,
- Shaping or changing music in a meaningful way, and
- Extracting meaning from music.

Many studies that elicit semantic descriptors of perceived changes in audio associated with the application of audio effects have been performed, as summarized in Table 1. In a series of papers, Brookes and Williams [4–6] investigated means to manipulate *brightness*, *warmth*, and *softness*. In [7], different audio effects in music production were assessed to determine how they can manipulate and influence perception of brightness and warmth. [8] developed an effect to independently control brightness and warmth. In [9], brightness and *sharpness* in distorted guitar content was explored. In [10] and references therein, the *punch* associated with a track was investigated, including the ability of dynamic range compression to make a track punchier. Similarly, [11] explored use of the semantic term *aggressive* in conjunction with a dynamic range compressor.

However, all of the above-mentioned studies dealt with one or two attributes in isolation. Another approach was presented in [12], which attempted to quantify a large number of semantic terms related to mix production and produced a table of frequency ranges and their associated semantic terms. However, this was based entirely on descriptions by practicing sound engineers, without any formal analysis that verified whether the descriptions were accurate or whether the terms were widely used.

Descriptive language is frequently used to refer to the way in which a mix or component in a mix is perceived [13]. Developing computational models of such timbral adjectives will advance knowledge of how the brain perceives audio signals and provide intuitive means for modeling production decisions and manipulating timbre.

In audio engineering, there is a corpus of terms that are widely accepted as having correlated spectral properties, referred to as a vocabulary of descriptors. Most experienced audio engineers and musicians tend to be in agreement regarding the perception of these descriptors, and such terms have formed the basis of audio effect presets across a wide range of software tools.

[14] provides a review and summary of a range of different audio descriptors commonly used to describe sound. Table 2 shows a selection of a list presented in [15]. It provides a review of spectral descriptors found in audio engineering literature, with their corresponding frequency ranges. Descriptors in this list are mostly associated with equalization effects and characterized by amplification or attenuation applied to a specific band in the magnitude spectrum of a sound.

As described in [16–19], descriptive language in music production can be categorized using a number of schemata. This allows for attribution of formal meaning to descriptions of sound, and separate potentially context-specific terms, such as those associated with an instrument, from terms that represent emotional or musically motivated responses.

Koelsch proposed a taxonomy for musical semantics in [20], in which subjective responses to musical sounds are grouped as having *extra-musical*, *intra-musical*, or *musicogenic* meaning. Extra-musical or *designative* meaning refers to associations between a musical sound source and non-musical context. Koelsch described three dimensions of extra-musical meaning: *iconic* referring to metaphorical comparisons between the sound and a non-musical quality (e.g., warm, sharp), *indexical* referring to the expression of an emotional state (e.g., joy, sadness), and *symbolic* referring to cultural and social references (e.g., a national anthem, relation of musical motives to an ethnic group).

Most of the semantic terms in music production fall into the *iconic* subcategory. This includes the schema from [21] for subjective responses to musical stimuli: *onomatopoeia* to represent terms that mimic the acoustic sound source, *sound source* to represent situational factors of the source (e.g., instrument, environment, etc.), and *adjectives* for a subjective, figurative description of the sound. A visual representation of this hierarchical taxonomy is presented in Fig. 1.

An alternative categorization of timbral adjectives is proposed in [22], which categorized terms based on the semantics of spatial audio processing [23, 24]: *technical*, referring

Table 1. An overview of studies of semantics describing audio effects. The terms column either lists the terms used in the study, for less than five terms, or the total number of unique terms used within the study. Some papers did not have associated specific terms but instead used gestural interfaces, so no terms were listed. Other terms used semantic web associates and, as such, could interrogate any number of terms. Some studies were review or combination studies, so they did not have participants or sample numbers to reference.

Audio Effect	Terms	Approach Project	No. Participants	No. Samples	Reference
Compression	976	Perceptual study SAFE	963	2,154	[30]
Compressor	Aggression, distortion	Perceptual study	17	2	[11]
Compressor	···	Gestural interface	20	1	[45]
Compressor	Rock, jazz, hiphop, EDM	Perceptual study	26	4	[46]
Compressor	Punch	Semantic rules to feature representation	···	···	[47]
Compressor	Clarity, punch	Signal analysis	8	2	[48]
Compressor	···	Semantic web feature analysis	···	···	[43]
Equalization	···	Review of semantic terms and frequency content	···	···	[49]
Equalization	681	Perceptual study SAFE	416	2,248	[30]
Equalization	Bright, warm	Perceptual study SAFE	59	1,113	[50]
Equalization	Bright, warm	Perceptual study SAFE	40	10	[32]
Equalization	324	Perceptual study SocialFX	633	3	[26]
Equalization	$\geq$6 billion	Machine learning word mapping	···	···	[51]
Gain	Balance	Perceptual study	25	5	[52]
Gain	Balance	Signal analysis	···	···	[53]
Reverb	747	Perceptual study SAFE	582	1,320	[30]
Reverb	3388	Perceptual study SocialFX	658	3	[28]
Reverb	9,161,912 dictionary	Semantic web mapping	···	265	[54]
Reverb	···	Semantic rules to feature representation	···	···	[55]
Spatialization	···	Semantic representation of metadata	···		[56]
Distortion	271	Perceptual study SAFE	135	444	[30]
Compressor, distortion, equalizer, reverb	618	Perceptual study SAFE	263	2,694	[2]
Compressor, distortion, equalizer, reverb	394	Combine study mapping SocialFX	432	4	[29]
Bitcrusher, distortion, compressor, equalizer	Bright, warm	Perceptual study	26	7	[7, 57]
Mixing systems	···	Semantic rule to audio feature representation	···	···	[58]
Mixing systems	···	Semantic web audio feature term modeling SAFE	···	···	[59]
Mixing systems	···	Review paper	···	···	[60, 44]
Mixing systems	···	Semantic web audio feature mapping	···	···	[39]

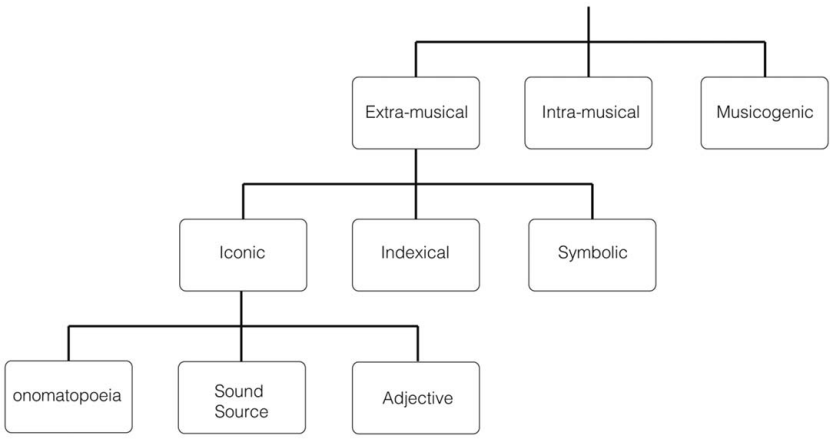


Fig. 1. The term classification hierarchy, proposed by [21] and [20].

to audio effect parameters of low-level audio features [25]; *spatial*, referring to position of a sound in an acoustic environment; or *timbral*, referring to perceptual dimension of timbre, often described using a metaphor.

1.1 Agreement Measures

The extent to which natural language can be used to provide novel methods of modifying sound is determined by the level of exhibited agreement within a term. When many test subjects use a term for a common purpose, it suggests there is a consensus on its perceptual representation. Terms with high consensus can be utilized to perform representative modifications of sounds using semantic interfaces.

In [13], participants were played randomized samples from the Freesound database[1] and asked to complete a survey attributing descriptive terms to each of the sounds. If the same sample was described using the same term by multiple participants, there must be a strong association between the two. *Bright*, *resonant*, and *harsh* all exhibited strong agreement scores, whereas *open*, *hard*, and *heavy* all showed low subjective agreement.

1.2 Similarity Measures

Determining the extent to which sound descriptions are similar allows for identification of synonyms (e.g., "Are bright and sharp timbrally equivalent?"), labelling of parameter scales (e.g., "Does something get warmer as it gets less bright?"), and recommendations for new settings based on current preferences (e.g., "It looks like you're trying to make this sound brighter. Try adjusting these parameters for better results."). Similarity measures can be divided into *Timbral* and *Contextual* Similarity.

2 THE DATA SETS

The three data sets analyzed herein are SocialFX, SAFE, and MEDS, each of which is described and outlined in this section.

2.1 SocialFX Data Set

The SocialFX database [1] contains the data from SocialEQ [26], Reverbalize [27], and an additional dynamic range compression study. The data set is very large, but because entries were crowdsourced using Amazon's Mechanical Turk platform, it is likely that only a small fraction of the contributors are expert users. Participants were removed from the study if too many default answers were submitted or if the participant was inconsistent, based on repeated tests.

Users contributed terms to the data set by positioning sounds in a reduced-dimensionality space, which corresponds to audio effect parameters. The top 50 SocialFX terms are presented inTable 3. Similar to the SAFE data set (SEC. 2.2), audio effect parameters are available for each instance, along with metadata extracted from the user.

[1] http://freesound.org.

Table 2. An excerpt from De Man [15].

Term		Range	Reference
Air		5–8 kHz	[61, p. 119]
		10–20 kHz	[62, p. 99]
		10–20 kHz	[63, p. 211]
		11–22.5 kHz	[61, p. 26]
		12–15 kHz	[64, p. 103]
		12–16 kHz	[65, p. 43]
		12–20 kHz	[62, p. 25]
		12–20 kHz	[66, p. 108]
		12–20 kHz	[67, p. 86]
Anemic	Lack of	20–110 Hz	[63, p. 211]
	Lack of	40–200 Hz	[61, p. 119]
Articulate		800–5,000 Hz	[61, p. 119]
Ballsy		40–200 Hz	[61, p. 119]
Barrelly		200–800 Hz	[61, p. 119]
Bathroomy		800–5,000 Hz	[61, p. 119]
Beefy		40–200 Hz	[61, p. 119]
Clarity		2.5–4 kHz	[67, p. 86]
		2.5–5 kHz	[68, p. 484]
		3–12 kHz	[63, p. 211]
		4–16 kHz	[61, p. 26]
Fat		50–250 Hz	[63, p. 211]
		60–250 Hz	[61, p. 25]
		62–125 Hz	[65, p. 43]
		200–800 Hz	[61, p. 119]
		240 Hz	[68, p. 484]
Presence		800–12,000 Hz	[61, p. 119]
		1.5–6 kHz	[62, p. 24]
		2–8 kHz	[65, p. 43]
		2–11 kHz	[63, p. 211]
		2.5–5 kHz	[68, p. 484]
		4–6 kHz	[61, p. 25]

The notion of inter-subject descriptor agreement was explored in [26], in which agreement for a descriptor d is represented by overall variance across participants for each of the dimensions in some statistical representation. For example, if all of the audio effect parameters from instances of users trying to make a sound *brighter* are observed, to what extent do each of those parameters vary?When taking the covariance matrix of these statistical parameters, this measurement is equivalent to the trace:

$$\text{trace}(\Sigma)_d = \frac{1}{N}\sum_{k=0}^{M}\sum_{n=0}^{N-1}(x_{n,k} - \mu_k)^2, \tag{1}$$

where N is the number of instances of the descriptor in the data set, M is the number of statistical parameters, $x_{n,k}$ is the k^{th} parameter for instance n, and μ_k is the mean of the k^{th} parameter across all instances of d. To measure agreement, the authors also take into consideration the number of entries made into the data set by dividing the natural log of the number of instances by the trace:

$$A_d = \frac{\ln N}{\text{trace}(\Sigma)_d}. \tag{2}$$

SocialEQ [26], Reverbalize [28, 27], and Audealize [29] took a similar approach to personalization of audio effect parameter spaces. Through repeated trials on example sounds, users taught the systems their own terms and parameter-space representations. Descriptive terms were

Table 3. The top 50 terms from the SocialFX [1] data set, sorted by number of instances.

N	Term	Total	Comp	EQ	Rev	N	Term	Total	Comp	EQ	Rev
1	Echo	2,396	118	0	2,278	26	The	397	22	1	374
2	Loud	1,308	261	21	1,026	27	Large	396	17	2	377
3	Tin	1,212	89	28	1,095	28	-Like	377	7	0	370
4	Low	1,154	92	16	1,046	29	And	361	21	10	330
5	War	1,137	147	60	930	30	Louder	350	156	0	194
6	Warm	1,057	135	59	863	31	Con	348	13	1	334
7	Church	1,033	8	0	1,025	32	Distorted	343	43	0	300
8	Big	934	55	1	878	33	Full	337	70	1	266
9	Spacious	855	62	0	793	34	Room	332	33	0	299
10	Distant	848	29	2	817	35	Nice	329	30	3	296
11	Deep	787	31	6	750	36	Drum	324	11	1	312
12	Muffle	634	85	4	545	37	Hollow	323	14	2	307
13	Muffled	623	81	4	538	38	Sad	323	3	21	299
14	Hall	584	7	0	577	39	High	319	37	4	278
15	Clear	567	126	8	433	40	Strong	316	40	1	275
16	Ring	537	24	7	506	41	Organ	315	0	0	315
17	Soft	533	102	26	405	42	Way	294	8	0	286
18	Big-	517	13	0	504	43	Pleasant	293	32	4	257
19	Bas	506	46	3	457	44	Under	286	7	1	278
20	Far	473	9	0	464	45	Old	282	18	36	228
21	Bass	461	43	1	417	46	Harp	277	55	8	214
22	Like	450	9	0	441	47	Smooth	277	13	9	255
23	Distort	442	62	0	380	48	Sound	277	31	1	245
24	Nic	432	34	6	392	49	Metal	270	23	2	245
25	Echoing	415	12	0	403	50	Sharp	257	55	7	195

Comp = compressor; Dist = distortion; EQ = equalizer; Rev = reverb.

then rendered in a 2D space, so, for example, users could control the amount of warmth or boominess of a signal.

2.2 The SAFE Data Set

Crowdsourced methods use templates from large data sets of descriptive terms to modify the timbre of a sound using audio effects. In [30, 2], terms are collected for equalization, dynamic range compression, distortion, and reverb. The data is sourced from users describing transformations that are made by the digital audio workstation plugins when applied to their own audio signals, within their own production workflow. Each term has a corresponding parameter set, audio feature set, and table of user metadata. Parameters can then be set based on average settings assigned to terms given by users. The system partitions the data, affording more specific presets based on external factors, such as the sound's genre and instrument.

This method was extended in [31] by allowing users to navigate the meaning of each term. Clustering was applied to a reduced dimensionality representation of the descriptor space, which was navigated using a machine learning autoencoder. This data has also been used to design semantically controlled audio effects [32], an audio processing chain generator [33], and a timbral modification [34].

In [2], parameter variance was used to cover the distribution of feature values for each term. Descriptors were mapped from various audio effect transforms into a common timbre space. The popularity of each descriptor was based on a coefficient representing the term as a proportion of the data set, Eq. (3).

$$p_d = c_d \times \ln \frac{n(d)}{\sum_{d=0}^{D-1} n(d)}, \tag{3}$$

where $n(d)$ is the number of entries for descriptor d, and c_d is the output of Eq. (1) when applied to the reduced dimensionality timbre space [2].

A comprehensive similarity matrix based on the term's context can be devised using techniques taken from natural language processing. In [2], a Vector Space Model was used to identify similarity of each term in a database, based on the number of entries from each audio effect. Fig. 2(a) shows the resulting pairwise similarities of the high-generality terms.

The most similar term pairs were *bass* and *strong*, *deep* and *sharp*, and *boom* and *thick*. Conversely, the similarity of transform types based on their descriptive attributes can be calculated by transposing the occurrence matrix in the Vector Space Model. Fig. 2(b) shows similar terms were used to describe equalization and distortion, whereas the equalization and compression vocabulary is more disjoint.

Timbral similarity between data points or clusters is often computed using audio features or audio effect parameters. One approach is to apply agglomerative clustering to a data set of labelled feature sets and then measure the cophenetic distance (the magnitude of the first common branch) between two data points. These cophenetic distances are shown in dendrogram plots, comparing the distances of terms for each audio effect in Fig. 3 using terms taken from the SAFE data set [30].

Table 4. The top 50 terms from the SAFE [30] data set, sorted by number of instances.

N	Term	Total	Comp	Dist	EQ	Rev	N	Term	Total	Comp	Dist	EQ	Rev
1	Warm	582	9	26	542	5	26	Gentle	11	6	2	1	2
2	Bright	531	4	5	521	1	27	Thick	11	2	2	6	1
3	Punch	34	27	1	6	0	28	Crushed	10	7	2	1	0
4	Room	33	1	0	2	30	29	Damp	10	1	1	1	7
5	Air	31	0	0	18	13	30	Harsh	10	1	4	5	0
6	Crunch	29	0	27	0	2	31	Low	10	0	0	10	0
7	Smooth	22	15	3	2	2	32	Presence	10	2	0	8	0
8	Vocal	22	16	1	4	1	33	Space	10	0	0	1	9
9	Clear	21	3	0	18	0	34	Tin	10	0	2	7	1
10	Subtle	21	6	4	1	10	35	Acoustic	9	4	2	3	0
11	Bass	20	3	4	13	0	36	Comp	9	9	0	0	0
12	Fuzz	19	1	17	1	0	37	Dream	9	1	0	0	8
13	Nice	18	12	0	4	2	38	Flat	9	5	1	3	0
14	Full	16	3	0	9	4	39	Hall	9	0	0	0	9
15	Boom	15	2	2	9	2	40	Kick	9	4	1	4	0
16	Crisp	15	1	3	11	0	41	Loud	9	6	2	1	0
17	Sofa	15	15	0	0	0	42	Present	9	3	0	6	0
18	Soft	15	5	1	4	5	43	Sharp	9	2	1	4	2
19	Big	13	1	0	1	11	44	Small	9	0	0	0	9
20	Clean	13	1	0	11	1	45	Bite	8	0	0	8	0
21	Thin	13	1	0	12	0	46	Click	8	1	0	7	0
22	Box	12	1	0	8	3	47	Cut	8	2	0	6	0
23	Deep	12	3	1	6	2	48	Dark	8	0	0	4	4
24	Tight	12	7	0	4	1	49	Echo	8	0	0	0	8
25	Drum	11	3	0	2	6	50	Glue	8	8	0	0	0

Comp = compressor; Dist = distortion; EQ = equalizer; Rev = reverb.

Resulting clusters are intended to retain perceived latent groupings, based on underlying semantic representations. In the EQ data [Fig. 3(c)], for example, terms associated with boosts in high-to-mid–frequency and high-frequency bands, such as *tin*, *cut*, *clear*, and *thin* are grouped together, whereas a cluster associated with boosts to low and low-to-mid bands are separated with high cophenetic distance. Fig. 4 shows that the spectral profiles of terms within the same cluster are highly correlated. Curves in the first cluster generally exhibit amplification around 500 Hz with a high-frequency roll-off. Similarly, terms in the second cluster

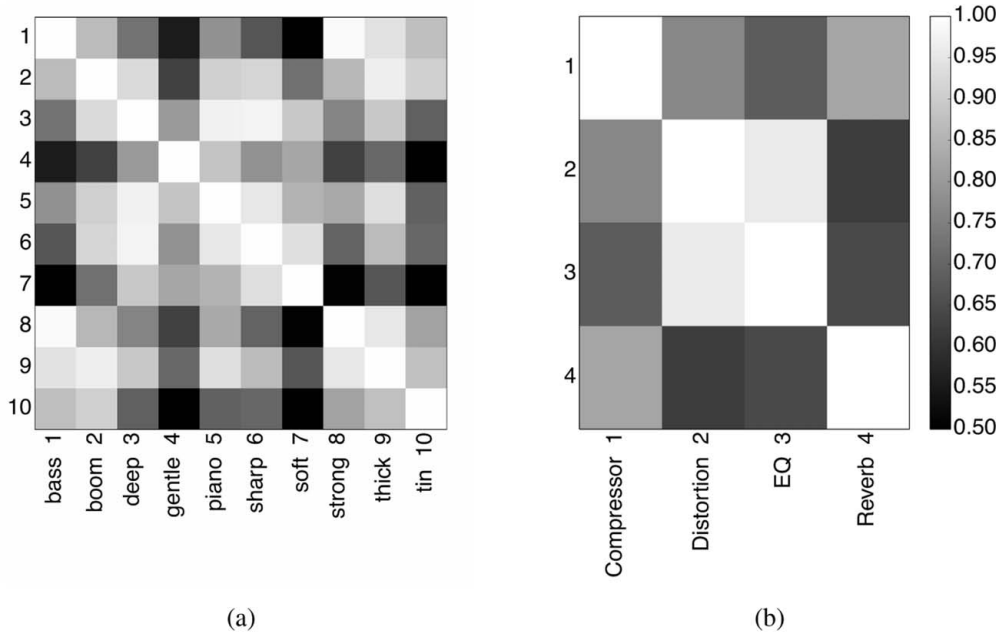


Fig. 2. Vector-space similarity with regard to (a) high-generality terms and (b) transform-classes. A cell with a high value (light shade) means that the horizontal term/transform is very similar to the vertical term/transform, with regard to the transforms/terms related to it. EQ = equalizer.

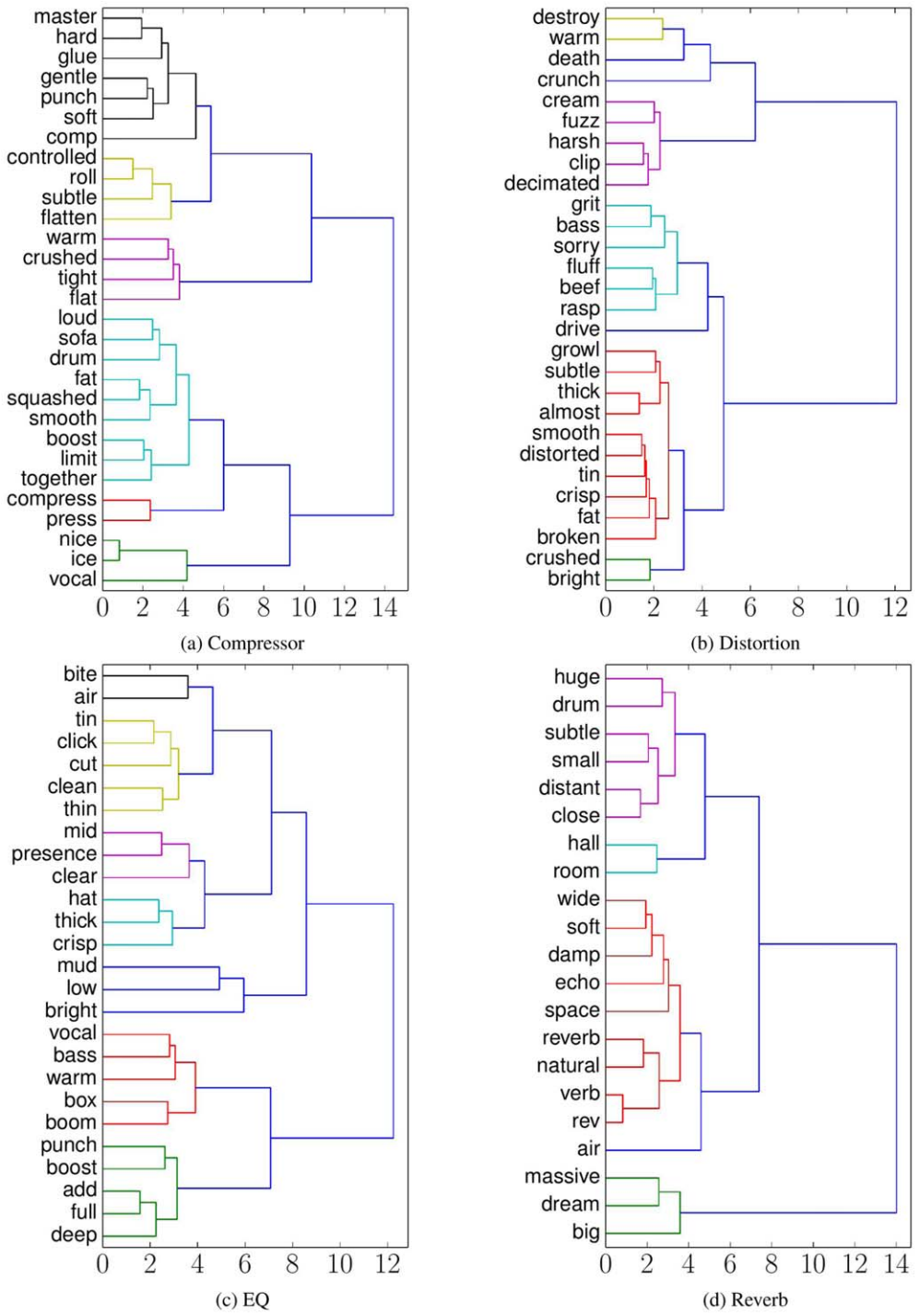
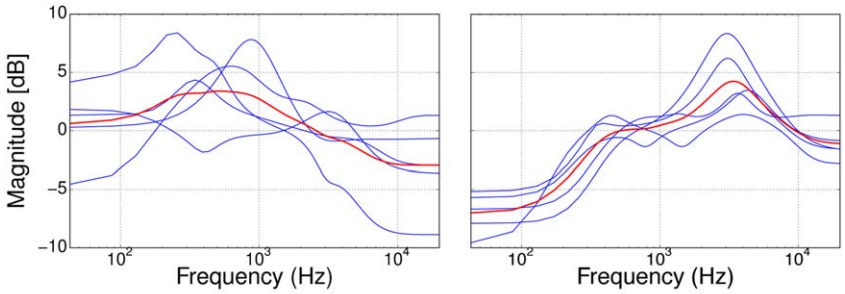


Fig. 3. Dendrograms showing term clustering based on feature space distances for each transform class (from [2]): (a) compressor, (b) distortion, (c) equalizer (EQ), and (d) reverb.

exhibit high frequency boost ($\geq$5 kHz) with attenuated low frequencies.

2.3 MEDS

The MEDS consists of mixes gathered in a real-life, ecologically valid setting and perceptual evaluation thereof, which can be used to expand knowledge on the mixing process. The data offers many opportunities for music production analysis, some of which are explored here. In particular, ratings and comments were collected by a wide range of listening test participants and subsequently annotated and analyzed. For instance, one finding was that more experienced subjects commented more on negative aspects (things that they thought could be improved) instead of on strengths, that they were more specific in their assessments of mixes, and that they were more likely to agree with one another.



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

Fig. 4. Two clusters of terms with similar equalization curves (thin lines) with their average (thick line). (a) Warm, bass, boom, box, and vocal. (b) Thin, clean, cut, click, and tin.

At the time of [3], MEDS contained around 20 multitrack recordings, most of them freely available. On average, there were ten mixes available for each of these songs, with the complete digital audio workstation session, so that the mix could be recreated and analyzed in detail [35]. Thus, [3] contains 180 mixes including parameter settings, close to 5,000 preference ratings [36] and free-form descriptions [12], and a diverse range of contributors from five different countries. It has since grown to include data from [37, 38] and an additional, unpublished study.

This allows for extensive research into the attributes of music production that are often lost when session information is discarded. By including deep metadata, the authors provide training data for fully automated music production systems [39].

Table 5. Number of audio samples per source data set.

Data Set	No. Samples
SocialFX	23,324
SAFE	10,347
MEDS	6,740

MEDS = Mix Evaluation Data Set.

Table 6. Number of audio samples per audio effect.

Audio Effect	No. Samples
Reverb	21,236
Compressor	6,847
Mix	6,740
Equalizer	4,854
Distortion	734

3 SECONDARY ANALYSIS

3.1 Combining Data Sets

These three data sets were combined in order to perform secondary analysis. To perform this analysis, the original data sets were collated, and audio feature analysis and extraction was performed.

Because each of the three data sets are related to transforming or manipulating audio and the corresponding semantic terms, the authors of this paper performed analysis of the data based on how each semantic descriptor transforms a set of audio features. This approach was inspired by [7].

The SAFE data set [30] relates a set of semantic terms to the corresponding effect parameter settings and audio features extracted both before and after the audio effect processing. The original audio content and audio effect processing was collected for both the MEDS [3] and SocialFX data sets [1]. In the case of MEDS, this audio transform was the full mixing process, so the original audio was defined as the direct sum of all audio stems. The purpose of this approach was to relate the associated semantic terms to the audio transformation. This approach was designed to reveal insight into how the audio changes, because semantic terms are all relative to the original audio content and context [40].

The combined three data sets consist of 40,411 audio samples labelled with 6,247 different descriptors, using five different audio effects. The number of individual audio samples present in each data set is presented in Table 5. The number of terms associated with each audio effect is presented in Table 6. With over 23,000 individual audio samples, the SocialFX data set contributes the largest proportion of the combined data set, more than double that of SAFE and over three times that of MEDS. Reverb is the most commonly used audio effect, and distortion is the least common, because this effect is only present in the SAFE data set. Mix represented the full mixing process with all possible audio effects included in this process.

Audio feature analysis was performed using the LibXtract library [41], in line with the SAFE data set. This process calculated 960 different audio features.

For further analysis of the semantic descriptors used, a Porter stemmer [42] was used to reduce each of the semantic terms to their core word, thereby removing differing endings while retaining the core semantic meaning of the word [30]. Based on this, a list of the most commonly occurring terms and their number of occurrences per audio effect and per data set can be seen in Tables 7 and 8, respectively. The list of stemmed words is limited to those that occur in all data sets for each comparison.

Table 7. The 21 most commonly used semantic descriptors and the number of occurrences of each term in each audio effect. Words are stemmed with suffixes stripped.

	Reverb	Compressor	Equalizer	Distortion	Mix	Total
Total	**21,236**	**6,847**	**4,854**	**734**	**6,740**	**40,411**
Warm	524	151	738	23	70	1,506
Bright	61	45	628	4	254	992
Loud	406	70	28	4	2	510
Muffl	294	48	6	1	130	479
Soft	167	50	30	5	108	360
Deep	260	16	15	4	14	309
Big	167	17	5	2	44	235
Harsh	46	9	24	4	134	217
Sharp	84	24	12	2	74	196
Strong	96	20	6	2	56	180
Smooth	100	34	17	6	14	171
Tinni	109	15	19	2	18	163
Metal	99	15	3	4	8	129
Cold	51	7	34	2	16	110
Vocal	17	49	29	2	2	99
Boomi	24	2	13	4	46	89
Crisp	42	25	10	2	10	89
Heavi	43	5	15	3	22	88
Hard	33	17	11	6	14	81
Punch	3	12	5	2	54	76
Subtl	14	23	1	1	14	53
Fuzzi	26	15	1	8	2	52
Kick	2	30	13	2	2	49

Table 7 shows that the term *warm* is one of the most occurring terms overall, also appearing as the highest occurring term in all effects except the mix. Terms such as *muffl* (from muffled) and *soft* are very prominent in the reverb, compressor, and mix terms. *Bright* seems to be particularly related to equalizer and mix but less relevant to reverb and compression when compared to other terms.

Table 8 demonstrates that the term *warm* is the most common term in both the SocialFX and SAFE data sets. *Bright* is the most commonly used term within the MEDS and the second-most term in the SAFE data set. Both *warm* and *bright* are the two most commonly used terms across all three data sets. The MEDS data set includes a number of terms that are less common in both SAFE and SocialFX, such as *thin*, *clear*, and *muddi*.

4 META-ANALYSIS

The three combined individual data sets were then evaluated further. To understand the impact the audio effect has on its raw audio input, more detailed analysis was performed. Following an approach similar to that used in [7], the set of audio features were extracted, both before and after the audio effect processing has taken place, using the LibXtract library [41]. This provides a delta audio feature representation of the audio effect processing, by taking the difference between the pre-audio and post-audio effect features. This delta audio feature representation is used because audio processing that takes place may depend heavily on the input audio signal and context [43, 39].

To better represent the large number of audio features in a smaller dimensional space, principal component analysis (PCA) was performed, reducing the dimensionality from 960 to 6 dimensions while still representing 97.6% of the variance in the data set, in an approach similar to [2]. From this point, the number of terms used was reduced to the 36 terms that occurred in all data sets, to ensure that each data set had representation across all terms and to remove single terms or sources of noise.

Table 8. The 36 most commonly used semantic descriptors and the number of occurrences of each term in each data set.

Term	MEDS	SAFE	SocialFX	Total
Warm	70	811	625	1,506
Bright	254	649	89	992
Loud	2	12	496	510
Muffl	130	1	348	479
Distant	18	11	436	465
Spaciou	6	4	427	437
Clear	194	21	179	394
Soft	108	34	218	360
Thin	318	10	16	344
Deep	14	22	273	309
Church	2	6	248	256
Dark	152	11	92	255
Big	44	18	173	235
Muddi	174	13	46	233
Distort	2	4	212	218
Harsh	134	14	69	217
Bass	4	49	153	206
Sharp	74	9	113	196
Dull	114	5	75	194
Strong	56	8	116	180
Full	36	22	121	179
Nice	2	28	147	177
Hollow	12	10	153	175
Smooth	14	35	122	171
Flat	126	5	33	164
Tinni	18	15	130	163
Cool	6	11	143	160
Mute	2	1	136	139
Quiet	20	6	107	133
Metal	8	14	107	129
Presenc	118	7	1	126
Live	12	5	103	120
Open	26	2	90	118
Close	42	4	67	113
Cold	16	6	88	110
Punchi	52	48	2	102

MEDS = Mix Evaluation Data Set.

Using this PCA feature representation, the effectiveness of this approach and similarity of semantic terms were calculated by finding the Ward linkage distance between each of the semantic terms, as proposed in [2]. The clustering of semantic terms can be seen in Fig. 5.

Upon visual inspection of the terms that are grouped together, it can be seen that there are a number of cases that are intuitive and make sense. The terms *tinni* and *metal* are grouped together, as are *big*, *full*, *strong*, and *hard*. Another group of *light*, *smooth*, *cold*, *quiet*, *crisp*, and *cool* makes intuitive sense, in terms of what would be expected from those semantic terms in a music production setting. Terms such as *bright*, *thin*, *warm*, and *room* tend to have less similarity to any other terms, but that could be because of the high number of occurrences of those terms in certain

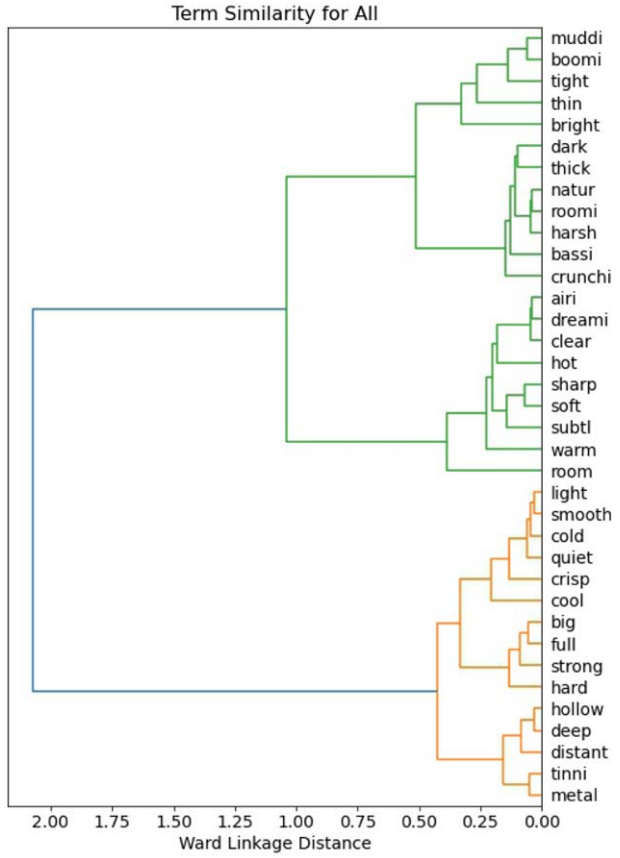


Fig. 5. The clustering of semantic terms across all data sets, calculated as the Ward linkage distance between principal component analysis (PCA) of the audio feature deltas.

subsets of the data, such that they dominate their own audio feature space.

To better investigate the agreement between each of these semantic terms, an agreement score is calculated for each term, using Eq. (1), taken from [26], as discussed in SEC. 2.1. This agreement score can be interpreted as the mean variance across all PCA audio feature dimensions that is normalized by the number of occurrences of the term. It is expected that higher occurrences of terms would produce a higher variance, so a higher agreement would be a large number of terms that produce a very low variance over the audio feature representation, whereas a low agreement would be a low number of terms that produce a higher variance in the audio feature domain. These agreement scores are presented across each data set in Fig. 6 and across each audio effect in Fig. 7, and in both cases, the scores are arranged in terms of highest to lowest agreement over the entire combined data set.

4.1 The Agreement of Semantic Terms

The agreement for each term in the combined data set was calculated with Eq. (2). This is then compared to the agreement score for each term within each individual data set. The results are presented in Fig. 6. These results show the highest agreement between terms such as *distant* and *deep*, and the results show that both of these terms have high agreement, particularly within the SocialFX data set. Both the terms *warm* and *dark* stand out when there is very high agreement of the term within the SocialFX data set, but when combined with other data sets, the overall agreement decreases. This demonstrates that some other data sets must disagree with the transformation that creates a warm or dark sound. In general, the SocialFX data set has the highest agreement across most terms, indicating that there is generally high agreement within this data set, whereas the MEDS data set generally has a lower agreement. Only in the cases of *thin* and *tight* is there higher agreement in the MEDS data set when compared with the SocialFX data set. The SocialFX data set also has higher agreement for every single term than the SAFE data set.

In general, SocialFX has a higher agreement score, and SAFE and MEDS are comparable on their agreement scores, with a few cases in which they differ considerably, such as *clear*, *strong*, *thin*, and *harsh*. This is most likely because of the SocialFX data set being constructed on just four different audio sample recordings, with individual audio effects applied, whereas the SAFE data set uses any audio sample the user wishes to select, the MEDS data set is constructed by a number of highly different songs, and the audio processing in a mix is considerably higher in complexity than a single audio effect.

The terms *warm* and *dark* both have very high agreement within the SocialFX data, but the combined agreement scores are lower. This would suggest that the MEDS and SAFE data sets disagree on the meaning of the terms when compared with SocialFX. For terms such as *distant* and *clear*, including the SAFE and MEDS data has negligible impact on the agreement score. This would suggest that all three data sets agree with the definition of these terms, in terms of the sonic transformation taking place.

Fig. 7 shows the agreement for each term in the combined data set, which was calculated with Eq. (2), compared with the agreement score for each term within each audio effect over the combined data source. This plot contains fewer terms, because only terms that existed in all five types of effects (compression, distortion, equalization, reverb, and mix) were included in this plot. The plot is sorted by the overall agreement when all audio effects are combined. In general, reverb consistently has the highest agreement score for all terms, sometimes giving a higher agreement than the combined audio effects, in the case of *deep* and *warm*. This could be because of the reverb effect having a considerably larger number of audio samples associated with it, at over 21,000 audio samples, which individually makes up over 52% of the total data set.

The equalizer audio effect has a high level of agreement for terms such as *soft*, *cold*, *warm*, and *bright*. Of these, *soft*, *warm*, and *bright* also have high levels of agreement for the mix effect. There is generally very low agreement for distortion terms, with the highest agreement being for the term *warm*, which also has the highest agreement for reverb effects and for the compressor. There is very high agreement with the terms *soft* for both the equalizer and compressor, and in both cases, these terms also have a

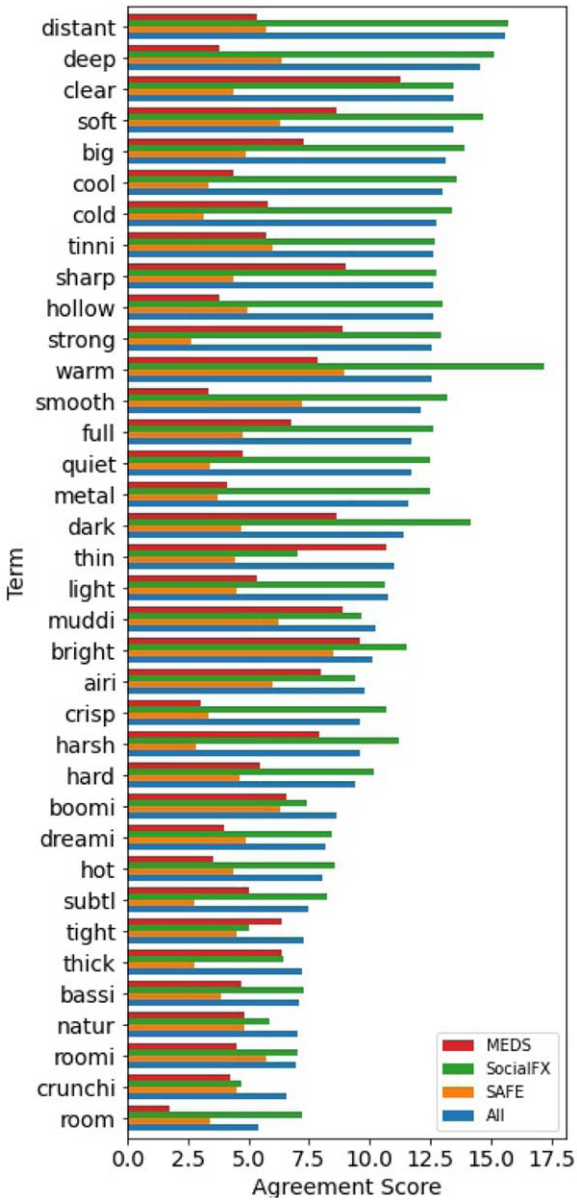


Fig. 6. The level of agreement of a term within the feature transform space, by data set, sorted by agreement score of all data sets combined, with highest agreement at the top. MEDS = Mix Evaluation Data Set.

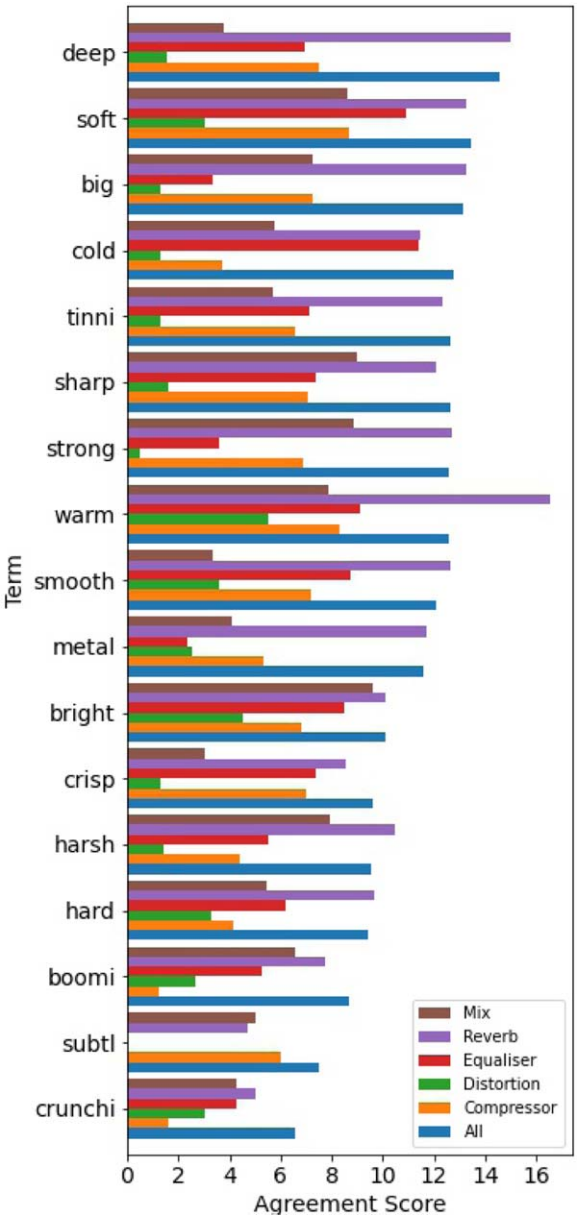


Fig. 7. The level of agreement of a term within the feature transform space, by effect, sorted by agreement score of all data sets combined, with highest agreement at the top.

high overall agreement, which suggests there is general agreement between the audio effects and within each audio effect individually.

Overall, for the term *bright*, each individual effect has a strong agreement score; however, when the audio effect data are combined, the agreement score does not increase. This suggests that the individual data sets agree with themselves but do not agree with data sets for other audio effects. A similar situation can be seen with terms such as *warm*, in which the reverb effect has a very high agreement score; although when combined with other audio effects, the combined agreement score decreases. This suggests that the audio effect used will impact the meaning of a semantic audio term, and the same term will have a different meaning when used in the context of a different audio effect.

4.2 Semantic Term Similarity

The audio feature PCA representation calculated per term, from SEC. 4, was used to calculate the mean average feature for each term. From this, a Pearson's correlation is performed to identify how similar the mean feature representation of each term is between the different data set pairs, which is shown in Fig. 8. Similarly, the mean average PCA audio feature representation of each term was calculated, and a Pearson's correlation was performed to identify the correlation between different audio effects with each individual term, which is shown in Fig. 9. These visualizations

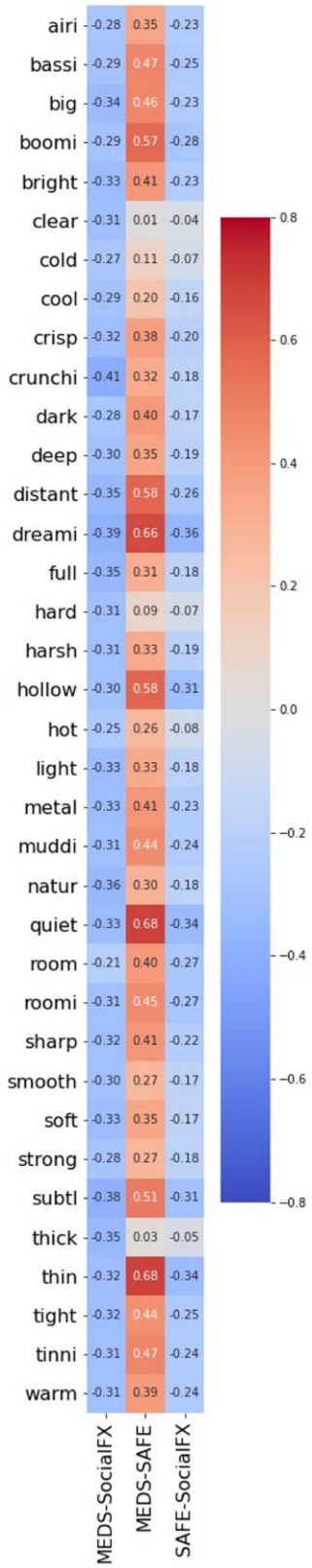


Fig. 8. Pearson's correlation between audio transformation of each semantic term, comparing each data set. The legend is shared with Fig. 9. MEDS = Mix Evaluation Data Set.

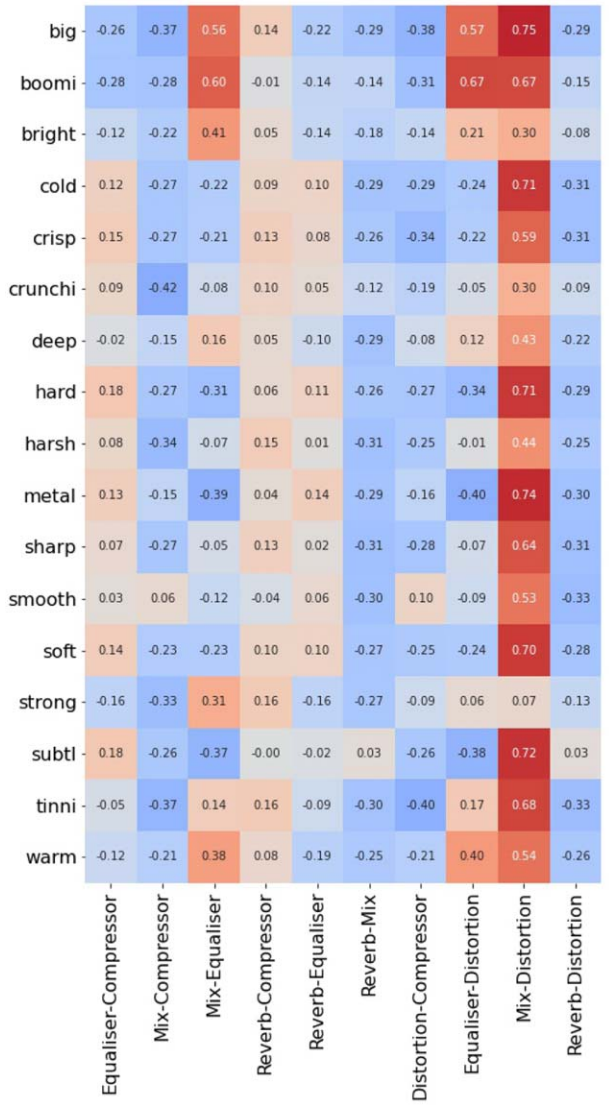


Fig. 9. Pearson's correlation between audio transformation of each semantic term, comparing each audio effect.

identify which semantic terms are similar across different data sets and audio effects.

Fig. 8 shows that there are no strong correlations between different data sets. Only the comparison between MEDS and SAFE data sets produce moderate correlations, where $abs(r) > 0.5$. The terms *boomi*, *distant*, *dreami*, *hollow*, *quiet*, *subtl*, and *thin* all produce moderately positive correlations, where $0.5 < abs(r) \leq 0.7$. In the case of these seven terms, both MEDS and SAFE reasonably agree on the audio transformation that takes place to describe these semantic terms. *Thin*, *quiet*, and *dreami* are the three highest correlations of terms. *Thin* also had the highest agreement score within the MEDS, in which the agreement score was higher than in SocialFX. Both *thick* and *hard* have almost no correlation between each other, and both had a relatively low agreement score too, so there is very little by way of agreement of terms for *thick* and *hard*.

The correlation between semantic terms by audio effect is visualized in Fig. 9. It can be seen that, in general, semantic terms associated with the mix effect and distortion effect correlate with each other quite strongly. Some of the terms have a high correlation [$abs(r) > 0.7$], and only the terms *bright*, *crunchi*, *deep*, *harsh*, and *strong* do not have at least a moderate correlation. The term *boomi* has a moderate correlation between mix and distortion, mix and equalizer, and distortion and equalizer, indicating that there is high agreement for this term between the three different audio effects. Other than that, the correlations between different audio effects are generally very low regarding the particular terms used. It can therefore be demonstrated that, with the exception of mix and distortion, individual semantic terms used will differ greatly depending on the audio effect that is used.

5 DISCUSSION

In this paper, three of the largest semantic music production data sets have been combined and compared to one another. The semantic terms are evaluated in terms of their direct modification to the audio attributes they represent.

The SocialFX data set tends to produce higher agreement scores than the MEDS and SAFE data sets. This may be because of some artefacts of the data set. In the creation of this data set, all semantic terms were evaluated on four audio samples, and audio effect ranges were limited in their modification to the source audio, (e.g., ± 4 dB on the equalizer). Furthermore, it is suspected that, in general, the SocialFX participants had less music production experience.

The SAFE data set produced a varied agreement over a range terms, with most agreement on the terms *warm* and *bright*. Within this study, the audio effect parameter ranges are much larger, and participants were given free reign to modify and control parameters, which resulted in a number of more extreme modifications to the audio content. Participants were also allowed to use any audio content they wanted and could submit any term they liked, which resulted in this data set being very noisy with large extremes in the data.

The MEDS data set represented the production and evaluation of music production mixes. The mix category is considerably more complex audio processing than any individual audio effect, and the parameters of the mix are still not fully understood by modern artificial intelligence mixing systems [44]. The MEDS data set used a smaller number of fixed tracks, with a large range of audio processing and thus permutations of the music production. The participants were all music production students or experts, so there is a guarantee of a benchmark quality, which cannot be said regarding the other data sets. Furthermore, it is not clear if a mono-sum of a series of audio tracks compared with a full mix-down is a reasonable comparison for an audio processing chain. This was used to represent the full mixing process; however, there may be alternative ways to review and improve this audio processing representation.

It has been identified that the SocialFX data set tends to have the highest agreement over all the audio effects and terms, which could be an artefact of the small musical variation size or because the experimental design was more strict than both SAFE and MEDS, which both gave participants a lot more independent control to modify parameters and add terms as they saw fit. There is higher correlation between terms in the MEDS and SAFE data sets, when compared with the SocialFX, and in audio effects, there is highest correlation between distortion and full-mix effects. It is suspected that the correlation between mix and distortion is because each of them only appear in one data set and thus represent the overall correlation between MEDS and SAFE.

Overall, there is high agreement on terms, such as "distant" and "clear," independent of the audio effect or data set used, with less agreement on well-known terms, such as "bright." It has also been shown that semantic terms vary greatly based on the audio effect being used to envisage the semantic term and that a "warm" equalizer will be very different from a "warm" distortion effect.

6 CONCLUSION

This article has presented an approach to collate a number of different data sets relating to music production mixing and audio effect semantics. The three largest known semantic data sets were collated and compared in a number of different ways to expose the commonalities and conflicts of these approaches.

It has been shown that each of the data sets can be combined to provide a greater insight into the semantics of music production. A variety of terms is used across a range of audio effects, and the limited agreements of terms across audio effects suggest that semantic terms vary depending on the audio effect used and the musical content and context of the piece being used.

Furthermore, the experimental designs of the different studies provide useful insights into the ways in which further studies on semantics can be conducted. Free-form studies in which participants select parameters and name them are very open but can result in very noisy results, which make is difficult to produce any considerable agreement, whereas highly specified studies with pre-selected parameters can force participants down a very specific path and have the danger of limiting the generalizability of the study. Clearly, there is a potential for some middle-ground approach, with a larger number of audio samples but limited set of terms than can apply, for future music production semantic research.

The results of this combined study demonstrate that semantic terms used to describe music production content are directly related to the type of processing applied. The type of audio processing will greatly impact the use of different semantic terms, and there is no universal semantic language for describing audio production, but different audio effects will produce different audio transformations that can independently be associated with given semantic terms.

7 ACKNOWLEDGMENT

The authors thank Bryan Pardo, Prem Seetharaman, Mark Cartwright, Zafar Rafii, Ryan Stables, and Sean Enderby for providing data.

8 REFERENCES

[1] T. Zheng, P. Seetharaman, and B. Pardo, "SocialFX: Studying a Crowdsourced Folksonomy of Audio Effects Terms," in *Proceedings of the 24th ACM International Conference on Multimedia*, pp. 182–186 (Amsterdam, The Netherlands) (2016 Oct.). https://doi.org/10.1145/2964284.2967207.

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

[3] B. De Man and J. D. Reiss, "The Mix Evaluation Dataset," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx)*, pp. 436–442 (Edinburgh, UK) (2017 Sep.).

[4] D. Williams and T. Brookes, "Perceptually-Motivated Audio Morphing: Softness," presented at the *126th Convention of the Audio Engineering Society* (2009 May), paper 7778.

[5] T. Brookes and D. Williams, "Perceptually-Motivated Audio Morphing: Brightness," presented at the *122nd Convention of the Audio Engineering Society* (2007 May), paper 7035.

[6] T. Brookes and D. Williams, "Perceptually-Motivated Audio Morphing: Warmth," presented at the *128th Convention of the Audio Engineering Society* (2010 May), paper 8019.

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

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

[9] K. Tsumoto, A. Marui, and T. Kamekawa, "The Effect of Harmonic Overtones in Relation to 'Sharpness' for Perception of Brightness of Distorted Guitar Timbre," *Proc. Mtgs. Acoust.*, vol. 29, no. 1, paper 035002 (2016 Nov.). https://doi.org/10.1121/2.0000380.

[10] S. Fenton, "Automatic Mixing of Multitrack Material Using Modified Loudness Models," presented at the *145th Convention of the Audio Engineering Society* (2018 Oct.), paper 10041.

[11] A. Moore, "Dynamic Range Compression and the Semantic Descriptor Aggressive," *Appl. Sci.*, vol. 10, no. 7, paper 2350 (2020 Mar.). https://doi.org/10.3390/app10072350.

[12] B. De Man and J. D. Reiss, "Analysis of Peer Reviews in Music Production," *JARP*, vol. 10 (2015 Jul.). https://www.arpjournal.com/asarpwp/analysis-of-peer-reviews-in-music-production/.

[13] M. Sarkar, B. Vercoe, and Y. Yang, "Words That Describe Timbre: A Study of Auditory Perception Through Language," in *Proceedings of the Language and Music as Cognitive Systems Conference*, pp. 37–38 (Cambridge, UK) (2007 May).

[14] A. Pearce, T. Brookes, and R. Mason, "Hierarchical Ontology of Timbral Semantic Descriptors," *Audio Commons: An Ecosystem for Creative Reuse of Audio Content,* D5.1 (2016 Aug.).

[15] B. De Man, *Towards a Better Understanding of Mix Engineering*, Ph.D. thesis, Queen Mary University of London, London, UK (2017 Jan.).

[16] A. Zacharakis, K. Pastiadis, J. D. Reiss, and G. Papadelis, "Analysis of Musical Timbre Semantics Through Metric and Non-Metric Data Reduction Techniques," in *Proceedings of the 12th International Conference on Music Perception and Cognition*, pp. 1177–1182 (Thessaloniki, Greece) (2012 Jul.).

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

[18] A. Zacharakis, K. Pastiadis, and J. D. Reiss, "An Interlanguage Unification of Musical Timbre," *Music Percept.*, vol. 32, no. 4, pp. 394–412 (2015 Apr.).

[19] A. Zacharakis, K. Pastiadis, G. Papadelis, and J. D. Reiss, "An Investigation of Musical Timbre: Uncovering Salient Semantic Descriptors and Perceptual Dimensions," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 807–812 (Miami, FL) (2011 Oct.).

[20] S. Koelsch, "Towards a Neural Basis of Processing Musical Semantics," *Phys. Life Rev.*, vol. 8, no. 2, pp. 89–105 (2011 Jun.). https://doi.org/10.1016/j.plrev.2011.04.004.

[21] S. Wake and T. Asahi, "Sound Retrieval With Intuitive Verbal Expressions," in *Proceedings of the International Community for Auditory Display (ICAD)* (Glasgow, UK) (1998 Nov.).

[22] E. R. Toulson, "A Need for Universal Definitions of Audio Terminologies and Improved Knowledge Transfer to the Audio Consumer," in *Proceedings of the 2nd Art of Record Production Conference* (Edinburgh, UK) (2006 Sep.).

[23] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.*, vol. 50, no. 9, pp. 651–666 (2002 Sep.).

[24] N. Ford, T. Nind, and F. Rumsey, "Communicating Listeners' Auditory Spatial Experiences: A Method for Developing a Descriptive Language," presented at the *118th*

Convention of the Audio Engineering Society (2005 May), paper 6481.

[25] D. Moffat, D. Ronan, and J. D. Reiss, "An Evaluation of Audio Feature Extraction Toolboxes," in *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx)*, pp. 277–283 (Trondheim, Norway) (2015 Nov.).

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

[27] P. Seetharaman and B. Pardo, "Crowdsourcing a Reverberation Descriptor Map," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 587–596 (New York, NY) (2014 Nov.). https://doi.org/10.1145/2647868.2654908.

[28] P. Seetharaman and B. Pardo, "Reverbalize: A Crowdsourced Reverberation Controller," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 739–740 (Orlando, FL) (2014 Nov.). https://doi.org/10.1145/2647868.2654876.

[29] P. Seetharaman and B. Pardo, "Audealize: Crowdsourced Audio Production Tools," *J. Audio Eng. Soc.*, vol. 64, no. 9, pp. 683–695 (2016 Sep.). https://doi.org/10.17743/jaes.2016.0037.

[30] R. Stables, S. Enderby, B. De Man, G. Fazekas, and J. D. Reiss, "SAFE: A System for the Extraction and Retrieval of Semantic Audio Descriptors," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)* (Taipei, Taiwan) (2014 Oct.), paper LBD15.

[31] S. Stasis, J. Hockman, and R. Stables, "Navigating Descriptive Sub-Representations of Musical Timbre," in *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*, pp. 56–61 (Copenhagen, Denmark) (2017 May).

[32] S. Stasis, R. Stables, and J. Hockman, "Semantically Controlled Adaptive Equalisation in Reduced Dimensionality Parameter Space," *Appl. Sci.*, vol. 6, no. 4, paper 116 (2016 Apr.). https://doi.org/10.3390/app6040116.

[33] S. Stasis, N. Jillings, S. Enderby, and R. Stables, "Audio Processing Chain Recommendation," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx)*, pp. 103–109 (Edinburgh, UK) (2017 Sep.).

[34] S. Enderby and R. Stables, "A Nonlinear Method for Manipulating Warmth and Brightness," in *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx)*, pp. 474–480 (Edinburgh, UK) (2017 Sep.).

[35] B. De Man, B. Leonard, R. King, and J. D. Reiss, "An Analysis and Evaluation of Audio Features for Multitrack Music Mixtures," in *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 137–142 (Taipei, Taiwan) (2014 Oct.).

[36] B. De Man, M. Boerum, B. Leonard, et al., "Perceptual Evaluation of Music Mixing Practices," presented at the *138th Convention of the Audio Engineering Society* (2015 May), paper 9235.

[37] A. Pras, B. De Man, and J. D. Reiss, "A Case Study of Cultural Influences on Mixing Practices," presented at the *144th Convention of the Audio Engineering Society* (2018 May), paper 9946.

[38] J. Colonel and J. D. Reiss, "Exploring Preference for Multitrack Mixes Using Statistical Analysis of MIR and Textual Features," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), e-Brief 526.

[39] D. Moffat, F. Thalmann, and M. B. Sandler, "Towards a Semantic Web Representation and Application of Audio Mixing Rules," in *Proceedings of the 4th Workshop on Intelligent Music Production (WIMP)* (Huddersfield, UK) (2018 Sep.).

[40] M. N. Lefford, G. Bromham, G. Fazekas, and D. Moffat, "Context-Aware Intelligent Mixing Systems," *J. Audio Eng. Soc.*, vol. 69, no. 3, pp. 128–141 (2021 Mar.). https://doi.org/10.17743/jaes.2020.0043.

[41] J. Bullock, "LibXtract: A Lightweight Library for Audio Feature Extraction," in *Proceedings of the International Computer Music Conference*, vol. 2007, pp. 25–28 (Copenhagen, Denmark) (2007 Aug.).

[42] M. F. Porter, "An Algorithm for Suffix Stripping," *Program: Electron. Libr. Inf. Syst.*, vol. 14, no. 3, pp. 130–137 (1980 Mar.).

[43] T. Wilmering, G. Fazekas, and M. B. Sandler, "High-Level Semantic Metadata for the Control of Multitrack Adaptive Digital Audio Effects," presented at the *133rd Convention of the Audio Engineering Society* (2012 Oct.), paper 8766.

[44] D. Moffat, "AI Music Mixing Systems," in E. R. Miranda (Ed.), *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity Music*, pp. 345–375 (Springer, Cham, Switzerland, 2021). https://doi.org/10.1007/978-3-030-72116-9_13.

[45] T. Wilson, S. Fenton, and M. Stephenson, "A Semantically Motivated Gestural Interface for the Control of a Dynamic Range Compressor," presented at the *138th Convention of the Audio Engineering Society* (2015 May), paper 9347.

[46] G. Bromham, D. Moffat, M. Barthet, and G. Fazekas, "The Impact of Compressor Ballistics on the Perceived Style of Music," presented at the *145th Convention of the Audio Engineering Society* (2018 Oct.), paper 10080.

[47] D. Moffat and M. B. Sandler, "Adaptive Ballistics Control of Dynamic Range Compression for Percussive Tracks," presented at the *145th Convention of the Audio Engineering Society* (2018 Oct.), e-Brief 484.

[48] S. Fenton, and J. Wakefield, "Objective Profiling of Perceived Punch and Clarity in Produced Music," presented at the *132nd Convention of the Audio Engineering Society* (2012 Apr.), paper 8626.

[49] B. De Man and J. D. Reiss, "A Semantic Approach to Autonomous Mixing," *JARP*, vol. 8 (2013 Dec.).

[50] S. Stasis, J. Hockman, and R. Stables, "Descriptor Sub-Representations in Semantic Equalization," in *Proceedings of the 2nd AES Workshop on Intelligent Music Production (WIMP2)* (2016 Sep.).

[51] S. Venkatesh, D. Moffat, and E. R. Miranda, "Word

Embeddings for Automatic Equalization in Audio Mixing," arXiv preprint arXiv:2202.08898 (2022 Feb).

[52] N. Jillings and R. Stables, "Investigating Music Production Using a Semantically Powered Digital Audio Workstation in the Browser," in *Proceedings of the AES International Conference on Semantic Audio* (2017 Jun.), paper P1-7.

[53] D. Moffat and M. B. Sandler, "Automatic Mixing Level Balancing Enhanced Through Source Interference Identification," presented at the *146th Convention of the Audio Engineering Society* (2019 Mar.), e-Brief 497.

[54] E. T. Chourdakis, J. D. Reiss, "Tagging and Retrieval of Room Impulse Responses Using Semantic Word Vectors and Perceptual Measures of Reverberation," presented at the *146th Convention of the Audio Engineering Society* (2019 Mar.), paper 10198.

[55] D. Moffat and M. B. Sandler, "An Automated Approach to the Application of Reverberation," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), paper 10264.

[56] A. Franck, J. Francombe, J. Woodcock, et al., "A System Architecture for Semantically Informed Rendering of Object-Based Audio," *J. Audio Eng. Soc.*, vol. 67, no. 7/8, pp. 498–509 (2019 Jul.). https://doi.org/10.17743/jaes.2019.0025.

[57] G. Bromham, D. Moffat, M. Barthet, and G. Fazekas, "Retro in Digital: Understanding the Semantics of Audio Effects," in *Proceedings of the Digital Music Research Network Workshop (DMRN)* (London, UK) (2019 Dec.).

[58] B. De Man and J. D. Reiss, "A Knowledge-Engineered Autonomous Mixing System," presented at the *135th Convention of the Audio Engineering Society* (2013 Oct.), paper 8961.

[59] S. Enderby, T. Wilmering, R. Stables, and G. Fazekas, "A Semantic Architecture for Knowledge Representation in the Digital Audio Workstation," in *Proceedings of the 2nd AES Workshop on Intelligent Music Production (WIMP2)* (2016 Sep.), paper 4102.

[60] D. Moffat and M. B. Sandler, "Approaches in Intelligent Music Production," *Arts*, vol. 8, no. 4, paper 125 (2019 Sep.). https://doi.org/10.3390/arts8040125.

[61] B. Owsinski, *The Mixing Engineer's Handbook* (Thomson Course Technology, Boston, MA, 2006), 2nd ed.

[62] K. Coryat, *Guerrilla Home Recording: How to Get Great Sound From Any Studio (No Matter How Weird or Cheap Your Gear Is)* (Hal Leonard Corporation, Milwaukee, WI, 2008), 2nd ed.

[63] R. Izhaki, *Mixing Audio: Concepts, Practices and Tools* (Focal Press, New York, NY, 2008), 1st ed.

[64] P. White, *Basic Mixers* (Sanctuary Publishing, London, UK, 1999).

[65] B. Katz, *Mastering Audio* (Focal Press, New York, NY, 2002).

[66] M. Cousins and R. Hepworth-Sawyer, *Practical Mastering: A Guide to Mastering in the Modern Studio* (Taylor & Francis, Milton Park, UK, 2013).

[67] G. Waddell, *Complete Audio Mastering: Practical Techniques* (McGraw-Hill Education, New York, NY, 2013).

[68] D. M. Huber and R. E. Runstein, *Modern Recording Techniques* (Focal Press, New York, NY, 2014), 8th ed.

THE AUTHORS


David Moffat


Brecht De Man


Joshua Reiss

David Moffat is an Artificial Intelligence and Machine Learning Research Scientist at Plymouth Marine Laboratory. He works on applying Artificial Intelligence and Machine Learning techniques to Earth Observation data to formulate a better understanding of the natural world. Previously he worked at Queen Mary University of London and the University of Plymouth as a post-doc and then academic, working on intelligent music production and music production semantics. He is the vice-chair of the AES Semantic Audio Analysis Technical Committee and a member of the AES UK committee.

•

Brecht De Man is Head of Research at PXL-Music, PXL University of Applied Sciences and Arts. He holds a Ph.D. in music mixing practices from the Centre for Digital Music at Queen Mary University of London, and he has presented, published, and patented research on intelligent audio effects, perception in sound engineering, and the analysis of music production practices. Brecht currently serves as Director of the AES.

•

Joshua Reiss is a Professor with the Centre for Digital Music at Queen Mary University of London. He has published more than 200 scientific papers and co-authored the book *Intelligent Music Production* and textbook *Audio Effects: Theory, Implementation and Application*. He is the President and a Fellow of the AES. He co-founded the highly successful spin-out company, LandR, and recently formed the start-ups Tonz and Nemisindo. His primary focus of research is on the use of state-of-the-art signal processing techniques for sound design and audio production.