Audio Engineering Society

Conference Paper

Presented at the Conference on
Semantic Audio
2017 June 22 – 24, Erlangen, Germany

Timbral attributes for sound effect library searching

Andy Pearce[1], Tim Brookes[1], and Russell Mason[1]

[1] *Institute of Sound Recording, University of Surrey, Guildford, Surrey, UK*

Correspondence should be addressed to Andy Pearce (andy.pearce@surrey.ac.uk)

ABSTRACT

To improve the search functionality of online sound effect libraries, timbral information could be extracted using perceptual models, and added as metadata, allowing users to filter results by timbral characteristics. This paper identifies the timbral attributes that end-users commonly search for, to indicate the attributes that might usefully be modelled for automatic metadata generation. A literature review revealed 1187 descriptors that were subsequently reduced to a hierarchy of 145 timbral attributes. This hierarchy covered the timbral characteristics of source types and modifiers including musical instruments, speech, environmental sounds, and sound recording and reproduction systems. A part-manual, part-automated comparison between the hierarchy and a *freesound.org* search history indicated that the timbral attributes *hardness*, *depth*, and *brightness* occur in searches most frequently.

1 Introduction

There are multiple online sound effect libraries that host sound effects available for use in audio production, such as *freesound.org*, *freeSFX.co.uk*, and *zapsplat.com*. Most of these libraries allow users to search for sound effects using keywords; finding sounds with matching titles and/or tags. Currently, tags are manually added by users, and are therefore non-standardised across all sound effects. Searches could be improved if all sounds had standardised tags related to characteristics such as timbre. It would be beneficial if these tags could be automatically generated, using perceptual models to predict timbral characteristics from features extracted from each audio file, as this would be quicker than manual tagging, and potentially more consistent.

If such functionality is to be developed then work should focus on the timbral attributes which users would find most useful (i.e. would search for most often). Therefore, this study has two aims: (*i*) to identify the attributes which can describe the timbral characteristics of sound effects; and (*ii*) to find the frequency-of-use for each of these attributes.

Many studies exist which have elicited timbral attributes; however, these studies are often focused on a specific type of sound (e.g. loudspeakers [1, 2, 3, 4], speech quality [5, 6, 7, 8], concert halls [9], etc.). In order to broaden the applicability of the current study's findings, a list of timbral attributes was first collated from a wide range of studies. The authors then structured these attributes into a hierarchy. This process is detailed in Section 2. Section 3 then describes how this hierarchy was used as a dictionary and compared against the search history from *freesound.org* (an online sound effect library which hosts Creative-Commons-licensed sound effects) to determine the frequency-of-use for each timbral attribute.

2 Attribute Identification

This section has three main aims: (1) identify timbral attributes from previous studies; (2) develop a *dictionary* of timbral terms, in a consistent adjectival format to compare against the search history; and (3) group and structure the timbral terms contained in the dictionary into a hierarchy of timbral attributes (with e.g. synonyms and antonyms grouped together).

2.1 Literature Attributes

In order make the list of timbral attributes as universal as possible, a wide range of published studies on timbral description was considered. Some of these focused on a particular stimulus type, such as environmental sounds, speech, musical instruments, concert halls, or sound recording and reproduction systems, though some covered multiple types.

In total, 1187 descriptors were identified, some of which were individual words and some of which were short phrases. The number of descriptors from each paper is shown in Table 1, along with the general topic of each paper. The full list of descriptors is included in the data repository available from doi: 10.5281/zenodo.167392.

Table 1: Number of descriptors from each source.

Source	Number of descriptors	Topic
Koivuniemi and Zacharov [10]	12	Spatial sound
Bagousse et al. [11]	28	Spatial sound
Barthet et al. [12]	47	Timbre ontology
Handel [13]	16	Psychoacoustics
Jensen [14]	8	Psychoacoustics
Zwicker and Fastl [15]	2	Psychoacoustics
Cano [16]	10	Sound description
Mattila [5, 6, 7, 8] (Summarised in [17])	27	Speech quality
Disley et al. [18]	17	Musical instruments
Wrzeciono and Marasek [19]	11	Musical instruments
Davies et al. [20]	49	Environmental/ Soundscape
Choisel [21]	8	Multichannel reproduced sound
Pedersen [22]	647	Reproduced sound
Pedersen and Zacharov [23]	42	Reproduced sound
Zacharov and Pedersen [24]	34	Reproduced sound
Gabrielsson and Sjögren [1]	13	Loudspeakers
Lavandier et al. [2]	3	Loudspeakers
Michaud et al. [3]	4	Loudspeakers
Staffeldt [4]	38	Loudspeakers
Lorho [25]	16	Headphones
Pearce et al. [26]	40	Microphones
Hermes [27]	105	Mix quality
Lokki et al. [9]	10	Concert halls

2.2 Attribute Reduction

Within the 1187 descriptors identified, there is likely to be a degree of redundancy, with multiple papers identifying the same descriptor or variations of it (e.g. brightness, bright, and brighter). Additionally, there may be descriptors that relate to aspects of sound that are not timbral (e.g. those to do with loudness, pitch, spatial, or musicological characteristics).

An automated removal of redundancy was followed by a manual removal of non-timbral attributes. To finally create the *dictionary* of terms, each descriptor was converted to an adjectival form, for example converting *noise* to *noisy*. This gave descriptors in a form likely be used as a timbral search term. For example, searches for *noise* will most likely not intend the word to be interpreted in its timbral sense (e.g. white noise), whereas searches for *noisy* will more likely be searches for sounds that have a noisy characteristic (e.g. noisy flute).

2.2.1 Automated redundancy removal

Redundancy in the data was automatically reduced by four natural language processing (NLP) methods: tokenizing, direct comparisons, lemmatization, and stemming, as in the work of Zacharov and Koivuniemi [28] and Guastavino and Katz [29]. Firstly, the 1187 descriptors were tokenized, which is an NLP expression indicating that each descriptor (which may include short phrases) was separated into its component words. This was conducted using the WordNet tokenizer package in Python 3.5 [30, 31].

Secondly, automated direct comparisons were made between all tokenized descriptors within the list, discarding any duplicates.

Thirdly, the remaining tokenized descriptors were *lemmatized* using WordNet. Lemmatization is a lexicographical transformation of a word to a common form. For example, the words "*transients*", "*transient's*", and "*transient*" would all be lemmatized to the word "*transient*". Lemmatization was followed by the removal of any duplicate lemmatized descriptors.

Finally, remaining descriptors were stemmed. Stemming is a cruder form of lemmatization, removing the suffixes of words to leave the base form of a word. For example, "*brightness*", "*brighter*", and "*brightest*" would all be stemmed to "*bright*". However, stemming can result in a word that is spelt incorrectly or has no meaning. For example, "*dense*", "*densest*" and "*denser*" will be stemmed to "*dens*". To prevent this, descriptors were stemmed and duplicates of the stemmed descriptors were removed, but an un-stemmed version of the descriptor was retained for the dictionary.

Using these four methods, the 1187 descriptors were reduced to 683.

2.2.2 Manual filtering

Following the automated redundancy removal, a manual approach was taken to remove non-timbral descriptors. This was completed via two tasks. Firstly, each of the three authors independently evaluated each of the 683 descriptors against the criterion:

> *A descriptor should be retained if it (or the adjectival form of it) describes a timbral characteristic of sound.*

Secondly, again independently, each author replaced each retained descriptor with its adjectival form. For example, "*depth*" was replaced by "*deep*". The adjectival forms are hereafter referred to as *terms*. The results were then compared across authors. Any terms deemed by all three to fail to meet the retention criterion were rejected. This left 224 terms that two or more authors agreed to retain, and 131 terms that only one author suggested retaining.

2.2.3 Group Discussion

A group discussion was held between the three authors to consider further each of the 355 retained terms. During this discussion, more detailed criteria for removing terms were developed and applied:

> *A term will be removed where it:*
>
> 1. *relates to loudness, pitch, or a spatial attribute;*
> 2. *refers to a musicological attribute;*
> 3. *is a hedonic or emotional term;*
> 4. *has meaning only with reference to another sound (real or imagined) that is not identified within the term (e.g. natural, realistic); or*
> 5. *can only refer to the relationship between a sequence of sounds.*

Where at least two authors agreed that a term failed one of the removal criteria, it was removed. This reduced the 335 terms to 295. These 295 terms form the *dictionary* that was later used for comparison against the *freesound.org* search history.

2.3 Timbral Attribute Grouping

Many of the terms within the dictionary relate to the same timbral attribute. For example, the terms "*bright*", "*dark*", and "*dull*" all relate to the timbral attribute of "*brightness*". To aid meaningful analysis of search frequencies, it is desirable to group the dictionary terms by timbral attributes. Additionally, it is desirable to structure these timbral attributes into a hierarchy as in the work of Pearce et al. [26] and Pedersen and Zacharov [23]. This has two benefits: (*i*) it allows for the frequency-of-use for terms that relate to the same perceptual attribute to be summed (e.g. summing the frequency-of-use for *bright*, *dark*, and *dull* to obtain

the frequency-of-use for the *brightness* attribute); and (*ii*) it allows for the frequency-of-use of each timbral attribute to be summed hierarchically (e.g. summing together the frequency-of-use for all attributes related to spectral balance).

The 295 terms within the dictionary were structured in this way by the three authors during a panel discussion. This resulted in 145 timbral attributes structured into a hierarchy, with 11 parent groups and up to four levels in each group. Interactive sunburst plots showing the structure of the hierarchy and the terms which comprise each attribute can be found at `http://iosr.uk/sunburst`. Alternatively, a high-resolution image of the hierarchy and the full list of timbral terms within each attribute can be found in doi: 10.5281/zenodo.167392.

2.4 Methodology discussion

The definition of timbre and its attributes is somewhat contentious [32, 33]. In order for the results of the current study to be as generalisable as possible, the authors have erred on side of exclusion, making it possible that some other researchers might feel that additional attributes could have been included, but less likely that any included attributes will be considered erroneous. Inclusion criteria have been intentionally strict, and both filtering and grouping have been repeated by an independent expert, whose findings were consistent with those of the authors.

3 Search Frequency

The frequency-of-use for each timbral term in the dictionary developed above was found using the search history of *freesound.org*, a popular online sound effect library which hosts Creative Commons licensed sound effects, with over 325,000 sound effects and over 4 million registered users.

3.1 Search Term Frequency

freesound.org retains the most recent month's search history. The analysis was conducted on the data for April, 2016. This provided a database of 8,154,586 searches (equivalent to 263,000 per day or 183 per minute), 879,976 of which were distinct. The data consisted of each distinct search, and its frequency (the number of times each distinct search was used for that month).

Each distinct search was tokenized using the WordNet tokenizer to split it into individual search words. Each search word was then compared against each dictionary term for an identical match. If a match was found, the frequency of the corresponding distinct search was added to the total for the matching dictionary term.

If no direct match was found, the similarity between each search word and each dictionary term was calculated using the WordNet Wu Palmer metric [34]: a measure of word similarity, ranging from 1.0 (perfect match) to 0.0 (no similarity). This metric is based on the distance between the two words within the WordNet taxonomy.

A threshold for the Wu Palmer similarity was set at 0.95, this value being determined by way of a trial-and-error manual optimisation process. If the similarity of a search word to a dictionary term was over 0.95 (i.e. a very high similarity), the frequency of the corresponding distinct search was added to the total for the matching dictionary term. For words which had multiple definitions within the WordNet taxonomy, the most common definition was used.

3.2 Manual Filtering

The dictionary term *screaming* was identified as the most frequently searched. However, closer inspection of the distinct searches in which this term occurred revealed that it was commonly being used not as a timbral descriptor (e.g. "screaming electric guitar tone") but as a verb (e.g. "woman screaming"). To remove the distinct searches where a dictionary term was not used as a timbral descriptor, the distinct searches were manually filtered.

There were 66,694 matches between distinct searches and dictionary terms. No automated method exists to determine if a search word is being used timbrally, and it was not practical to manually inspect all 66,694 matching distinct searches; instead a combination of two more efficient manual filtering methods was employed: term-specific filtering, and overall filtering.

3.2.1 Term-Specific Manual Filtering

For each dictionary term, the 50 most frequently used matching distinct searches, a total of 8615 searches, were manually inspected to give an indication of the proportion of distinct searches which were not using the term timbrally. This task was completed by the three authors, with the instructions:

Include a distinct search only if the term is used unambiguously as the intended timbral descriptor, indicated by the hierarchical grouping.

Ambiguity can result from:

- *The word being used in isolation (e.g. "screaming");*
- *The word being potentially used as a verb (e.g. "woman screaming");*
- *The word being potentially used as a noun (e.g. "female scream"); or*
- *The word being used as an adjective meaning something different from what our hierarchy intends (e.g. "noisy children").*

This manual filtering removed 7111 distinct searches. The expected proportion of timbre-related searches for each dictionary term was then obtained by dividing the frequency-of-use for each dictionary term's retained searches by the total frequency-of-use for the dictionary term's analysed searches. This proportion was then applied to the total frequency-of-use for each dictionary term to give the weighted frequency-of-use.

This weighted frequency-of-use for each dictionary term was then summed according to the attribute grouping discussed in Section 2.3. The 40 most searched timbral attributes are shown in Figure 1, along with the cumulative distribution for these attributes.

Only the most frequently used matching distinct searches were inspected, rather than a random sample of all matching distinct searches, since this represents a much larger proportion of matching searches overall. However, the generalisability of any method inspecting only a subset of distinct matching searches can not be guaranteed. As a validity check, a different (although not necessarily entirely independent), non-term-specific, subset was also inspected. Broad agreement across the two subsets would provide at least an indication of likely generalisability.

3.2.2 Overall Manual Filtering

Across all dictionary terms, the matching distinct searches were ranked by their frequency-of-use. Then, the 10,000 most frequently used distinct searches were taken for analysis. Any matching distinct searches that contained only a single word (as this met the exclusion criteria described in Section 3.2.1) or were removed from the term-specific manual filtering were removed. This left 6,617 distinct searches.

These were inspected by the three authors using the exclusion criteria set out in Section 3.2.1. The frequency of the non-excluded distinct searches was then summed for each dictionary term and, as with the previous analysis, the frequency-of-use for each dictionary term was then summed to identify the frequency of each timbral attribute. Each timbral attribute's frequency-of-use using this analysis method is shown in Figure 2, along with the cumulative distribution.

3.3 Comparing Frequencies of Timbral Attributes

The Spearman's correlation coefficient between the rank orders from the two methods is 0.779 ($p < 0.001$). Although this indicates that there is similarity between the two methods, there is some difference in the rank order. Comparing the rank order of the 40 most frequently searched attributes across both methods shows high rank correlation ($\rho = 0.935, p < 0.001$).

By visually inspecting Figures 1 and 2, it can be seen that the three most frequently searched timbral attributes are identical: *hardness*, *depth*, and *brightness*. However, the fourth and fifth attributes, *electronic-nature* and *weight*, are interchanged. The attribute of *swoosh*, ranked as 13th in the term-specific filtering method, was ranked as 6th in the overall filtering method.

As can be seen in both Figures 1 and 2, the frequency-of-use for timbral attributes diminishes very quickly with the rank order. This indicates that the majority of timbral searches are for the highest ranked few timbral attributes.

4 Summary

This paper had two aims: (*i*) to identify the attributes which can describe the timbral characteristics of sound effects; and (*ii*) to find the frequency-of-use for each of these attributes when users search online sound effect libraries.

To meet aim *i*, timbral descriptors were collated and parsed from multiple literature sources to create a *dictionary* of 295 timbral terms. These terms were

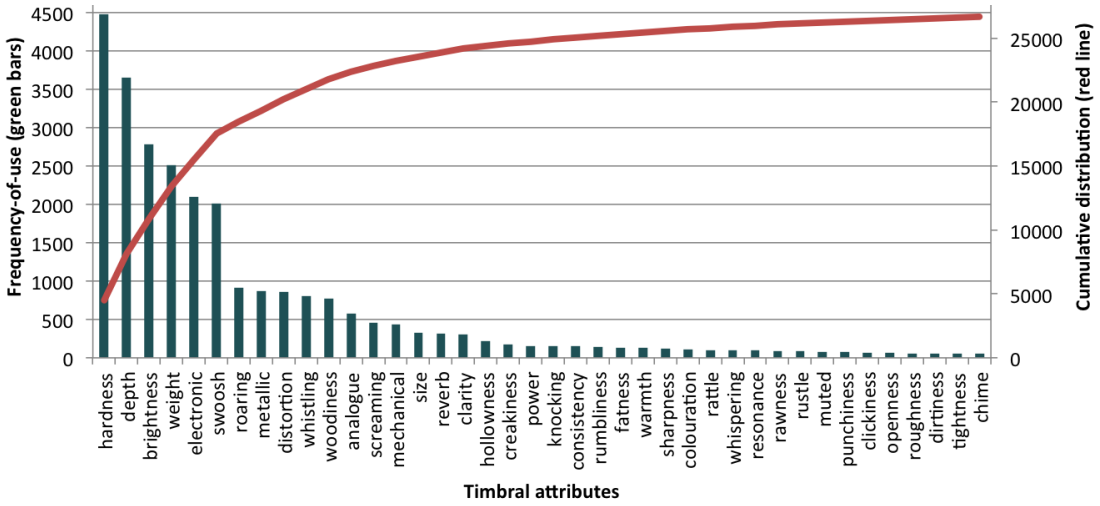

Fig. 1: Weighted frequency of use and cumulative distribution for the 40 most frequently searched timbral attributes based on the term-specific filtering method.

grouped into a hierarchy of 145 timbral attributes. This hierarchy covered the timbral characteristics of source types and modifiers including musical instruments, speech, environmental sound sources, concert halls and sound recording and reproduction systems.

Aim *ii* was met by comparing this dictionary against one month's search history from *freesound.org* with two manual filtering methods to ensure that matching searches used the terms as timbral descriptors.

Comparisons across both methods revealed that the terms *hardness*, *depth*, and *brightness* were the most searched for attributes. The fourth and fifth most frequently searched attributes differed between the analysis methods. The term-specific manual filtering identified the attributes of *electronic-nature* and *swoosh* as fourth and fifth most searched respectively, whereas the overall manual filtering identified the *weight* and *electronic-nature* attributes.

The results of this study provide an indication of the attributes that might usefully be modelled for automatic generation of timbral metadata for use in audio search engines. They also have the potential to feed into ongoing research into semantic feature extraction [35] and similarity-based recommendation [36].

5 Acknowledgements

This research was completed as part of the AudioCommons research project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688382. The data underlying the findings presented in this paper are available from doi: 10.5281/zenodo.167392. Further project information can be found at `http://www.audiocommons.org`.

References

[1] Gabrielsson, A. and Sjögren, H., "Perceived sound quality of sound-reproducing systems," *J. Acoust. Soc. Am.*, 65(4), pp. 1019–1033, 1979.

[2] Lavandier, M., Meunier, S., and Herzog, P., "Identification of some perceptual dimensions underlying loudspeaker dissimilarities," *J. Acoust. Soc. Am.*, 123(6), pp. 4186–4198, 2008.

[3] Michaud, P., Lavandier, M., Meunier, S., and Herzog, P., "Objective characterization of perceptual dimensions underlying the sound reproduction of 37 single loudspeakers in a room," *Acta Acustica with Acustica*, 101, pp. 603–615, 2015.

[4] Staffeldt, H., "Correlation between subjective and objective data for quality loudspeakers," in *47th Convention of the Audio Eng. Soc.*, Copenhagen, Denmark, 1974.

[5] Mattalia, V., "Descriptive analysis of speech quality in mobile communications: descriptive lan-

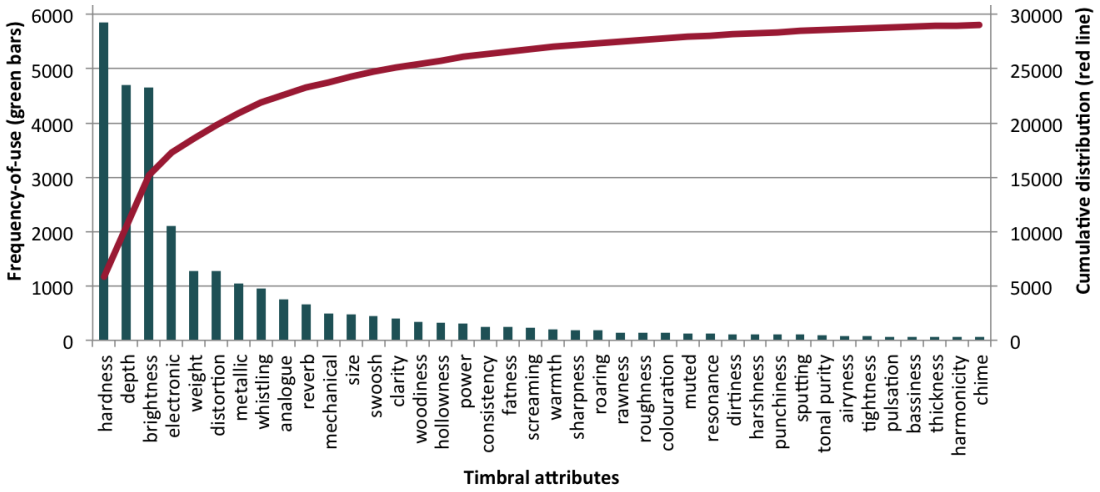

Fig. 2: Frequency-of-use and cumulative distribution for the 40 most frequently searched timbral attributes based on the 10,000 most frequently used matching distinct searches.

guage development and external preference mapping," in *111th Convention of the Audio Eng. Soc.*, New York, USA, 2001.

[6] Mattila, V., *Perceptual analysis of speech quality in mobile communications*, Ph.D. thesis, Tampere University of Technology, Tampere, Finland, 2001.

[7] Mattila, V., "Descriptive analysis and ideal point modelling of speech quality in mobile communications," in *113th Convention of the Audio Eng. Soc.*, Los Angeles, USA, 2002.

[8] Mattila, V., "Semantic analysis of speech quality in mobile communications: descriptive language development and mapping to acceptability," *Food quality and preference*, 14, pp. 441–453, 2003.

[9] Lokki, T., Pätynen, J., Kuusinen, A., Vertanen, H., and Tervo, S., "Concert hall acoustics assessment with individually elicited attributes," *J. Acoust. Soc. Am.*, 130(2), pp. 835–849, 2011.

[10] Koivuniemi, K. and Zacharov, N., "Unraveling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training," in *111th Convention of the Audio Eng. Soc.*, New York, USA, 2001.

[11] Bagousse, S., Paquier, M., and Colomes, C., "Families of sound attributes for assessment of spatial audio," in *129th Convention of the Audio Eng. Soc.*, San Francisco, USA, 2010.

[12] Barthet, M., Fazekas, G., Juric, D., Pauwels, J., Sandler, M., and Vetter, L., "Deliverable D2.1 - Requirements report and use cases," Available: `http://www.audiocommons.org/materials/`, AudioCommons project, 2016.

[13] Handel, S., "Timbre perception and auditory object identification," in B. Moore, editor, *Hearing*, chapter 12, pp. 425–461, Academic Press, San Diago, CA, 1995.

[14] Jensen, K., "The timbre model," Available: `http://www.musanim.com/pdf/JensenTimbreModel.pdf`, University of Copenhagen, Music informatics laboratory, ND.

[15] Zwicker, E. and Fastl, H., *Psycho-acoustics: Facts and models*, Springer, Berlin, Germany, 2nd edition, 2007.

[16] Cano, P., *Content-based audio search: from fingerprinting to semantic audio retrieval*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2007.

[17] Bech, S. and Zacharov, N., *Perceptual Audio Evaluation: Theory, Method and Application*, Wiley, West Sussex, England, 2006.

[18] Disley, A., Howard, D., and Hunt, A., "Timbral description of musical instruments," in *Proceedings of 9th International Conference on Music Perception and Cognition*, pp. 61–68, Bologna, Italy, 2006.

[19] Wrzeciono, P. and Marasek, K., "Violin sound quality: expert judgements and objective measures," in Z. Raś and A. Wieczorkowska, editors, *Advances in Music Information Retrieval*, chapter 3, pp. 237–260, Springer-Verlag Berlin Heidelberg, Berlin, Germany, 2010.

[20] Davies, W., Adams, M., Bruce, N., Cain, R., Carlyle, A., Cusack, P., Hall, D., Hume, K., Irwin, A., Jennings, P., Marselle, M., Place, C., and Poxon, J., "Perception of soundscapes: An interdisciplinary approach," *Applied Acoustics*, (74), pp. 224–231, 2013.

[21] Choisel, S., "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," *J. Acoust. Soc. Am.*, 121(1), pp. 388–400, 2007.

[22] Pedersen, T., "The semantic space of sounds: Lexicon of sound-describing words," Available: `http://assets.madebydelta.com/docs/share/Akustik/The_Semantic_Space_of_Sounds.pdf`, Delta Labs, 2008.

[23] Pedersen, T. and Zacharov, N., "The development of a sound wheel for reproduced sound," in *138th Convention of the Audio Eng. Soc.*, Warsaw, Poland, 2015.

[24] Zacharov, N. and Pedersen, T., "Spatial sound attributes: development of a common lexicon," in *139th Convention of the Audio Eng. Soc.*, New York, USA, 2015.

[25] Lorho, G., "Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating," in *118th Convention of the Audio Eng. Soc.*, Barcelona, Spain, 2005.

[26] Pearce, A., Brookes, T., Mason, R., , and Dewhirst, M., "Eliciting the most prominent perceived differences between microphones," *J. Acoust. Soc. Am.*, 139(5), pp. 2970–2981, 2016.

[27] Hermes, K., "Towards measuring and modelling the perceived quality of music mixes," Phd confirmation report, University of Surrey, 2014.

[28] Zacharov, N. and Koivuniemi, K., "Audio descriptive analysis and mapping of spatial sound displays," in *Proceedings of the 2001 International Conference on Audiatory Display*, Espoo, Finland, 2001.

[29] Guastavino, C. and Katz, B., "Perceptual evaluation of multi-dimensional spatial audio reproduction," *J. Acoust. Soc. Am.*, 116, pp. 1105–1115, 2004.

[30] Bird, S. and Loper, E., "NLTK 3.0 documentation: nltk.tokenize package," 2016.

[31] Miller, G., "WordNet: A Lexical Database for English," *Communications of the ACM*, 38(11), pp. 39–41, 1995.

[32] Hajda, J., Kendall, R., Carterette, E., and Harshberger, M., "Methodological issues in timbre research," in I. Deliége and J. Sloboda, editors, *Perception and cognition of music*, pp. 253–306, Psychology Press, New York, NY, 1997.

[33] Krumhansl, C., "Why is musical timbre so hard to understand?" in S. Nielzén and O. Olsson, editors, *Structure and perception of electroacoustic sound and music*, volume 846, pp. 43–53, Excerpta Medica, 1989.

[34] Wu, Z. and Palmer, M., "Verb semantics and lexical selection," *32nd annual meeting of the Association of Computational Linguistics*, pp. 133–138, 1994.

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

[36] Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, E., A. Gómez, and Herrera, P., "Semantic audio content-based music recommendation and visualisation based on user preference examples," *Information Processing & management*, 49(1), pp. 13–33, 2013.