UNSUPERVISED TAXONOMY OF SOUND EFFECTS

David Moffat
Centre for Digital Music,
Queen Mary University of London
London, UK
d.j.moffat@qmul.ac.uk

David Ronan
Centre for Intelligent Sensing,
Queen Mary University of London
London, UK
d.m.ronan@qmul.ac.uk

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

ABSTRACT

Sound effect libraries are commonly used by sound designers in a range of industries. Taxonomies exist for the classification of sounds into groups based on subjective similarity, sound source or common environmental context. However, these taxonomies are not standardised, and no taxonomy based purely on the sonic properties of audio exists. We present a method using feature selection, unsupervised learning and hierarchical clustering to develop an unsupervised taxonomy of sound effects based entirely on the sonic properties of the audio within a sound effect library. The unsupervised taxonomy is then related back to the perceived meaning of the relevant audio features.

1. INTRODUCTION

Sound designers regularly use sound effect libraries to design audio scenes, layering different sounds in order to realise a design aesthetic. For example, numerous explosion audio samples are often combined to create an effect with the desired weight of impact. A large part of this work involves the use of foley, where an artist will perform sound with a range of props. A key aspect of foley is that the prop being used may not match the object in the visual scene, but is capable of mimicking its sonic properties. An example would be the use of a mechanical whisk, which becomes a convincing gun rattle sound effect when combined in a scene with explosions and shouting.

Sound designers are less interested in the physical properties or causes of a sound, and more interested in their sonic properties. Despite this, many sound effect libraries are organised into geographical or physical categories. In [1] a sound search tool based on sonic properties is proposed, considering loudness, pitch and timbral attributes. A similar tool for semantic browsing of a small library of urban environmental sounds has also been proposed [2]. No other known classification methods for sound effects based on their sonic attributes exist, instead most previous work focuses either on perceptual similarity or the context and source of the sound.

Given that the practical use for a sound sample is often abstracted from its original intention, source or semantic label, categorisation based on this information is not always desirable. Furthermore, no standard exists for the labelling of recorded sound, and the metadata within a sound effect library can be highly inconsistent. This makes the task of searching and identifying useful sounds extremely laborious, and sound designers will often resort to recording new sound effects for each new project.

The aim of this paper is to produce a hierarchical taxonomy of sound effects, based entirely on the sonic properties of the audio samples, through the use of unsupervised learning. Such an approach offers an alternative to standard categorisation, in the hope that it will aid the search process by alleviating dependence on hand written labels and inconsistent grouping of sounds.

Different approaches to developing taxonomies of sound are discussed in Section 2. Section 3 presents the dataset, feature selection technique and unsupervised learning method undertaken to produce a hierarchy within a sound effect library. The taxonomy we produced is presented in Section 4. The evaluation of the presented taxonomy is undertaken in Section 4.4 and discussed in Section 5. Finally, the validity of the taxonomy and future work is discussed in Section 6.

2. BACKGROUND

There are a number of examples of work attempting to create a taxonomy of sound. In [3], the author classified sounds by acoustics, psychoacoustics, semantics, aesthetics and referential properties. In [4], the authors classified "noise-sound" into six groups: roars, hisses, whispers, impactful noises, voiced sounds and screams. This is further discussed in [5].

Production of a taxonomy of sounds heard in a cafe or restaurant were produced, basing the grouping on the sound source or context [6, 7].

In [8] the authors presented a classification scheme of sounds based on the state of the physical property of the material. The sound classifications were vibrating solids, liquids and aerodynamic sounds (gas). A series of sub-classifications based on hybrid sounds were also produced along with a set of properties that would impact the perception of the sound. This was developed further by attempting to understand how participants would arbitrarily categorise sounds [9]. In [10] the authors asked participants to identify how similar sounds are to each other along a series of different dimensions. They then performed hierarchical cluster analysis on the results, to produce a hierarchical linkage structure of the sounds. Furthermore, in [11] the authors performed a similar study where participants were asked how alike sets of sounds were. Audio features were then correlated to a likeness measure and a hierarchical cluster was produced on the set of selected features.

In [12] the authors asked participants to rate the similarity of audio samples, and performed hierarchical cluster analysis to demonstrate the related similarity structure of the sounds. Acoustic properties of sound walk recordings were taken and unsupervised clustering performed in [13]. These clusters were identified and related back to some semantic terms. Similarly, sound walks and interviews were used to identify appropriate words as sound descriptors [14]. Classification of sound effects by asking individuals to identify suitable adjectives to differentiate different sound samples was performed in [15] and similarly in [16] where the authors define classes of sound descriptor words that can be used to

Reference	Type of Sound	Classification properties	Quantitive Analysis	Qualitative Analysis	Word Classification	Audio Feature Analysis	Hierarchical Cluster
[3]	Environmental	Acoustics	N	N	N	Y	N
[3]	Environmental	Aesthetics	N	N	N	N	N
[3]	Environmental	Source/context	N	N	N	N	Y
[4, 5]	Environmental	Subjective	N	N	N	N	N
[6]	Cafe sounds	Source or context	N	N	N	N	Y
[7]	Restaurant	Subjective 'liking' score	N	Y	Y	N	N
[7]	Restaurant	Word occurrence	N	Y	Y	N	Y
[8]	Environmental	Physical properties	Y	N	N	N	N
[9]	Environmental	Subjective grouping	Y	N	N	N	Y
[10]	Environmental	Subjective ratings	Y	N	N	Y	Y
[11]	Environmental	Subjective ratings	Y	N	N	N	Y
[12]	Environmental	Subjective ratings	Y	N	Y	N	Y
[13]	Sound walks	Low level audio features	Y	Y	N	Y	N
[14]	Sound walks	Semantic words	Y	N	Y	N	N
[15]	Soundscape	Subjective free text word recurrence	N	Y	Y	N	N
[16]	'Perceptual attribute' words	Definition of word	N	Y	Y	N	N
[17]	Broadcast objects	Predefined word list	Y	Y	Y	N	Y
[18]	Urban sounds	Source	N	N	N	N	Y
[19]	Synthesised sounds	Control parameters	N	N	N	N	Y
[20]	Field recordings	Labels/audio features	Y	N	N	Y	N

Table 1: *Summary of literature on sound classification*

relate the similarity of words. In an extension to this, [17] asked participants to perform a sorting and labelling task on broadcast audio objects, again yielding a hierarchical cluster.

[18] produced a dataset of urban sounds, and a taxonomy for the dataset, where sounds are clustered based on the cause of the audio, rather than the relative similarity of the audio sample themselves. They then used this dataset for unsupervised learning classification [21, 22]. In the context of synthesised sounds, [19] grouped sounds by their control parameters.

There is no clear standard method for grouping sounds such as those found in a sound effect library. It becomes clear from the literature that there is limited work utilising audio features to produce a taxonomy of sound. It can be seen in Table 1 that a large range of relevant work structures sound grouping based on either subjective rating or word clustering. It is also apparent there is little work clustering the acoustic properties of individual samples. There is a discussion of sound classification based on the acoustic properties of samples [3], but only a high level discussion is presented and is not pursued further.

3. METHODOLOGY

We used unsupervised machine learning techniques to develop an inherent taxonomy of sound effects. This section will detail the various development stages of the taxonomy, as presented in Figure 1. The Adobe sound effects library was used. A set of audio features were extracted, feature selection that was performed using Random Forests and a Gaussian Mixture Model was used to predict the optimal number of clusters in the final taxonomy. From the reduced feature set, unsupervised hierarchical clustering was performed to produced the number of clusters as predicted using the Gaussian Mixture Model. Finally the hierarchical clustering results are interpreted. All software is available online [1].

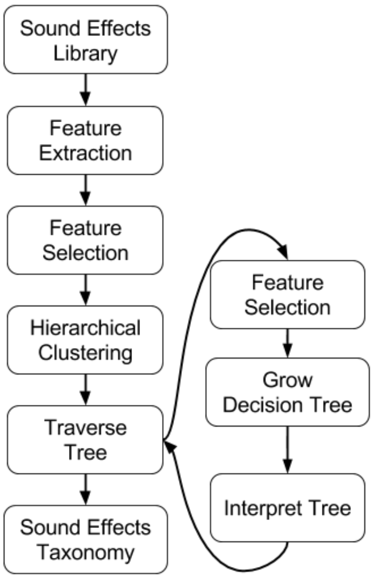

Figure 1: *Flow Diagram of unsupervised sound effects taxonomy system.*

3.1. Dataset

A dataset containing around 9,000 audio samples from the Adobe sound effect library [2] is used. This sound effects library contains a range of audio samples. All input audio signals were downmixed to mono, downsampled to 44.1 kHz if required, and had the initial and final silence removed. All audio samples were loudness normalised using ReplayGain [23]. Each sound effect was placed in a different folder, describing the context of the original sound effect. The original labels from the sound effect library can be found in Table 2, along with the number of samples found in each folder.

[1] `https://goo.gl/9aWhTX`

[2] `https://goo.gl/TzQgsB`

Class Name	Quantity of Samples	Class Name	Quantity of Samples
Ambience	92	Animals	173
Cartoon	261	Crashes	266
DC	6	DTMF	26
Drones	75	Emergency Effects	158
Fire and Explosions	106	Foley	702
Foley Footsteps	56	Horror	221
Household	556	Human Elements	506
Impacts	575	Industry	378
Liquid-Water	254	Multichannel	98
Multimedia	1223	Noise	43
Production Elements	1308	Science Fiction	312
Sports	319	Technology	219
Tones	33	Transportation	460
Underwater	73	Weapons	424
Weather	54		

Table 2: *Original label classification of the Adobe Sound Effects Dataset. DC are single DC offset component signals. DTMF is Dual Tone Multi Frequency - a set of old telephone tones.*

3.2. Feature Extraction

The dataset described in Section 3.1 was used. We used Essentia Freesound Extractor to extract audio features [24], as Essentia allows for extraction of a large number of audio features, is easy to use in a number of different systems and produced the data in a highly usable format [25]. 180 different audio features were extracted, and all frame based features were calculated using a frame size of 2048 samples with a hop size of 1024, with the exception of pitch based features, which used a frame size of 4096 and the hop size 2048. The statistics of these audio features were then calculated, to summarise frame based features over the audio file. The statistics used are the mean, variance, skewness, kurtosis, median, mean of the derivative, the mean of the second derivative, the variance of the derivative, the variance of the second derivative, the maximum and minimum values. This produced a set of 1450 features, extracted from each file. Sets of features were removed if they provided no variance over the dataset, thus reducing the original feature set to 1364 features. All features were then normalised to the range $[0, 1]$.

3.3. Feature Selection

We performed feature selection using a similar method to the one described in [26], where the authors used a Random Forest classifier to determine audio feature importance.

Random forests are an unsupervised classification technique where a series of decision trees are created, each with a random subset of features. The out-of-bag (OOB) error was then calculated, as a measure of the random forests classification accuracy. From this, it is possible to allocate each feature with a Feature Importance Index (FII), which ranks all audio features in terms of importance by evaluating the OOB error for each tree grown with a given feature, to the overall OOB error [27].

In [26] the authors eliminated the audio features from a Random Forest that had an FII less than the average FII and then grew a new Random Forest with the reduced audio feature set. This elimination process would repeat until the OOB error for a newly grown Random Forest started to increase.

Here, we opted to eliminate the 1% worst performing audio features on each step of growing a Random Forest, similar to but more conservative than the approach in [28]. In order to select the correct set of audio features that fit our dataset we chose the feature set that provided us with lowest mean OOB error over all the feature selection iterations.

On each step of the audio feature selection process, we cluster the data using a Gaussian Mixture Model (GMM). GMM's are an unsupervised method for clustering data, on the assumption that data points can be modelled by a gaussian. In this method, we specify the number of clusters and get a measure of GMM quality using the Akaike Information Criterion (AIC). The AIC is a measure of the relative quality of a statistical model for a given dataset. We keep increasing the number of clusters used to create each GMM, while performing 10-fold cross-validation until the AIC measure stops increasing. This gives us the optimal number of clusters to fit the dataset.

3.4. Hierarchical Clustering

There are two main methods for hierarchical clustering. Agglomerative clustering is a bottom up approach, where the algorithm starts with singular clusters and recursively merges two or more of the most similar clusters. Divisive clustering is a top down approach, where the data is recursively separated out into a fixed number of smaller clusters.

Agglomerative clustering was used in this paper, as it is frequently applied to problems within this field [10, 11, 12, 13, 26, 17]. It also provides the benefit of providing cophonetic distances between different clusters, so that the relative distances between nodes of the hierarchy are clear. Agglomerative clustering was performed, on the feature reduced dataset, by assigning each individual sample in the dataset as a cluster. The distance was then calculated for every cluster pair based on Ward's method [29],

$$d(c_i, c_j) = \sqrt{\frac{2n_{c_i}n_{c_j}}{n_{c_i} + n_{c_j}}} euc(x_{c_i}, x_{c_j}) \tag{1}$$

where for clusters c_i and c_j, x_c is the centroid of a cluster c, n_c is the number of elements in a cluster c and $euc(x_{c_i}, x_{c_j})$ is the euclidean distance between the centroids of clusters c_i and c_j. This introduces a penalty for clusters that are too large, which reduces the chances of a single cluster containing the majority of the dataset and that an even distribution across a hierarchical structure is produced. The distance is calculated for all pairs of clusters, and the two clusters with the minimum distance d are merged into a single cluster. This is performed iteratively until we have a single cluster. This provides us with a full structure of our data, and we can visualise our data from the whole dataset, down to each individual component sample.

3.5. Node Semantic Context

In order to interpret the dendrogram produced from the previous step, it is important to have an understanding of what is causing the separation at each of the node points within the dendrogram. Visualising the results of machine learning algorithms is a challenging task. According to [30] decision trees are the only classification method which provides a semantic explanation of the classification. This is because a decision tree faciliates inspection of individual features and threshold values, allowing interpretation of the separation of different clusters. This is not possible with any other classification methods. As such, we undertook feature selection and then grew a decision tree to provide some semantic meaning to the results.

Each node point can be addressed as a binary classification problem. For each node point, every cluster that falls underneath one side is put into a single cluster, and everything that falls on the other side of the node is placed in another separate cluster. Everything that does not fall underneath the node is ignored. This produces two clusters, which represent the binary selection problem at that node point. From this node point, a random forest is grown to perform the binary classification between the two sets and feature selection is then performed as described in Section 3.3. The main difference here is that only the five most relevant features, based on the FII are selected at each stage.

A decision tree is grown with this reduced set of 5 audio features, to allow manual visualisation of the separation of data at each node point within the hierarchical structure. The decision tree is constructed by minimising the Gini Diversity Index (GDI), at each node point within the decision tree, which is calculated as:

$$GDI = 1 - \sum_{i} p(i)^2 \tag{2}$$

where i is the class and $p(i)$ is the fraction of objects within class i following the branch. The decision trees are grown using the CART algorithm [31]. To allow for a more meaningful visualisation of the proposed taxonomy, the audio features and values were translated to a semantically meaningful context based on the audio interpretation of the audio feature. The definitions of the particular audio features were investigated and the authors identified the perceptual context of these features, providing relevant semantic terms in order to describe the classification of sounds at each node point.

4. RESULTS AND EVALUATION

4.1. Feature Extraction Results

Figure 2 plots the mean OOB error for each Random Forest that is grown for each iteration of the audio feature selection process. In total there were 325 iterations of the feature selection process, where the lowest OOB error occurred at iteration 203 with a value of 0.3242. This reduced the number of audio features from 1450 to 193.

Figure 3 depicts the mean OOB error for each Random Forest feature selection iteration against the optimal amount of clusters, where the optimal amount of clusters was calculated using the AIC for each GMM created. We found the optimal amount of clusters to be 9, as this coincides with the minimum mean OOB error in Figure 3.

4.2. Hierarchical Clustering Results

Having applied agglomerative hierarchical clustering to the reduced dataset, the resultant dendrogram can be seen in Figure 4. The dotted line represents the cut-off for depth analysis, chosen based on the result that the optimal choice of clusters is 9.

The results of the pruned decision trees are presented in Figure 5. Each node point identified the specific audio feature provides the best split in the data, to create the structure as presented in Figure 4.

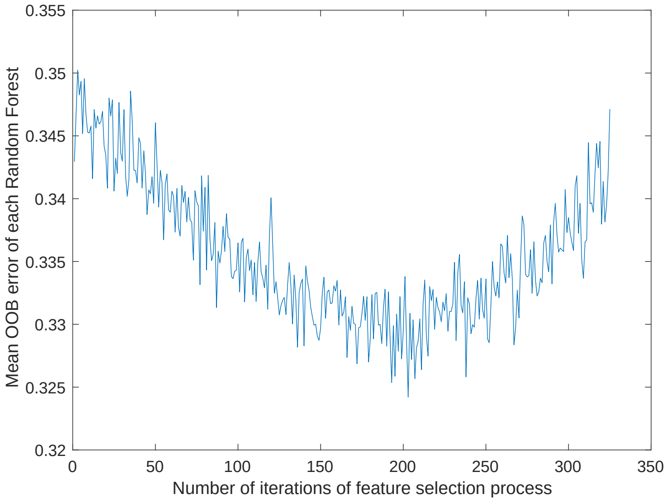

Figure 2: *Mean OOB Error for each Random Forest grown plotted against number of feature selection iterations*

4.3. Sound Effects Taxonomy Result

The audio features used for classification were related to their semantic meanings by manual inspection of the audio features used and the feature definitions. This is presented in Figure 6. As can be seen, the two key factors that make a difference to the clustering are periodicity and dynamic range.

Periodicity is calculated as the relative weight of the tallest peak in the beat histogram. Therefore strongly periodic signals have a much higher relative peak weight than random signals, which we expect to have near-flat beat histograms. Dynamic range is represented by the ratio of analysis frames under 60dB to the number over 60dB as all audio samples were loudness normalised and all leading and trailing silence was removed, as discussed in Section 3.2. Further down the taxonomy, it is clear that periodicity stands out as a key factor, in many different locations, along with the metric structure of periodicity, calculated as the weight of the second most prominent peak in the beat histogram. Structured music with beats and bars will have a high metrical structure, whereas single impulse beats or ticks will have a high beat histogram at one point but the rest of the histogram should look flat.

4.4. Evaluation

To evaluate the results of the produced sound effect taxonomy, as presented in Figure 6, the generated taxonomy was compared to the original sound effect library classification scheme, as presented in Section 3.1. The purpose of this is to produce a better understanding of the resulting classifications, and how it compares to more traditional sound effects library classifications. It is not expected that out clusters will appropriately represent an pre-existing data clusters, but that it may give us a better insight into the representation of the data.

Each of the 9 clusters identified in Figures 5 and 6 were evaluated by comparing the original classification labels found in Table 2 to the new classification structure. This is presented in Figure 7, where each cluster has a pie chart representing the distribution of original labels from the dataset. Only labels that make up more than 5% of the dataset were plotted.

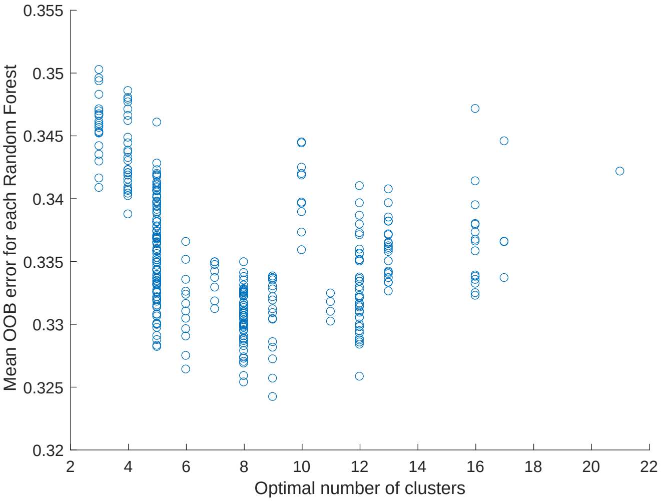

Figure 3: *Mean OOB Error for each Random Forest grown plotted against optimal number of clusters for each feature selection iteration*

In cluster 1, which has quick, periodic, high dynamic range sounds with a gradual decay, the majority of the results are from a range of production elements which are highly reverberant repetitive sounds, such as slide transition sounds. Many of these sounds are artificial or reverberant in nature, which follows the intuition of the cluster identification.

Cluster 2 contains a combination of foley sounds and water-splashing sounds. These sounds are somewhat periodic, such as lapping water, but do not have the same decay as in cluster 1.

Cluster 3 is very mixed. Impacts, household sounds and foley make up the largest parts of the dataset, but there is also contribution from crashes, production elements and weapon sounds. It is clear from the distribution of sounds that this cluster contains mostly impactful sounds. It is also evident that a range of impactful sounds from across the sound effect library have been grouped together.

In cluster 4, most of the samples are from the production elements label. These elements are moderately periodic at a high rate, such as clicking and whooshing elements, which are also similar to the next category of multimedia.

Cluster 5 contains a spread of sound labels, which includes transport and production elements as the two largest components. In particular, the transport sounds will be a periodic repetition of engine noises or vehicles passing, while remaining at a consistent volume.

There is a large range of labels within cluster 6. The three most prominent are human, multimedia and production elements, though cartoon and emergency sounds also contribute to this cluster. Human elements are primarily speech sounds, so the idea that periodic sounds that do not have a lot of high mid seems suitable, as the human voice fundamental frequency is usually between 90Hz and 300Hz.

Cluster 7 is entirely represented by the science fiction label. These fairly repetitive, constant volume sounds have an unnaturally large amount of high mid frequency.

Within cluster 8, the largest group of samples is multimedia, which consists of whooshes and swipe sounds. These are aperiodic, and their artificial nature suggests a long reverb tail or echo. A low dynamic range suggests that the samples are consistent in loudness, with very few transients.

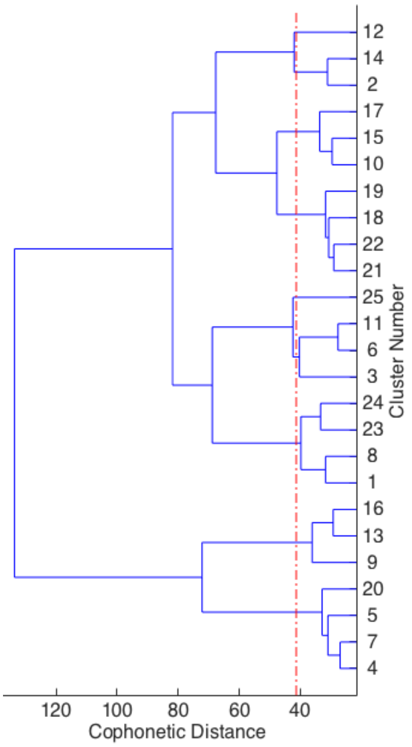

Figure 4: *Dendrogram of arbitrary clusters - The dotted line represents the cut-off for the depth of analysis (9 clusters)*

Finally, cluster 9 consists of a range of aperiodic impactful sounds from the impact, foley, multimedia and weapon categories.

5. DISCUSSION

The 9 inferred clusters were compared to the 29 original labels. It is clear that some clusters relate to intuition, and that this structure may aid a sound designer and present a suitable method for finding sounds, such as impactful sounds in cluster 9. Despite this, there are some clusters that do not make intuitive sense, or are difficult to fully interpret. We suspect that this is due to the depth of analysis on the dataset. Despite the GMM predicting 9 clusters within the data, we believe that a greater depth of analysis and clustering could aid in providing more meaningful, interpretable results, as many of the clusters are currently too large.

As can be seen from Figure 6 and discussed in Section 4, dynamic range and periodic structure are the key factors that separate this dataset. It is surprising that no timbral attributes and only one spectral attribute appears in the top features for classification within the dataset, and that seven of the eight features are time domain features.

Cluster 7 was described entirely as ‘Science Fiction’ in Section 4.4. This set of sound effects is entirely artificial, created using synthesisers and audio production. We believe that that the grouping using this audio feature is an artefact of the artificial nature of the samples and the fact they all come from a single source. This

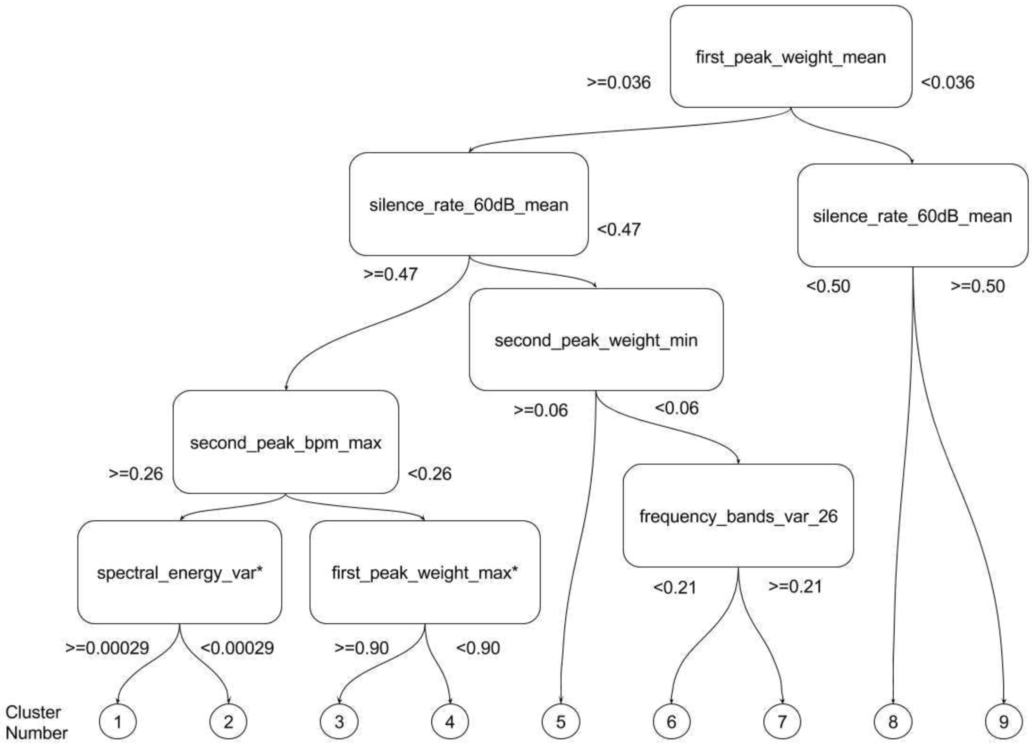

Figure 5: *Machine learned structure of sound effects library, where clusters are hierarchical clusters. The single audio feature contributing to the separation is used as the node point,with normalised audio feature values down each branch to understand the impact the audio feature has on the sound classification. The $*$ represents a feature separation where the classification accuracy is less than 80%, never less than 75%.*

is also caused by the analysis and evaluation of a single produced sound effect library. This artefact may be avoided with a large range of sound effects from different sources.

Section 4.4 shows that the current classification system for sound effects may not be ideal, especially since expert sound designers often know what sonic attributes they wish to obtain. This is one of the reasons that audio search tools have become so prominent, yet many audio search tools only work using tag metadata and not the sonic attributes of the audio files.

Our produced taxonomy is very different from current work. As presented in Section 2, most literature bases a taxonomy on either audio source, environmental context or subjective ratings.

6. CONCLUSION

Given a commercial sound effect library, a taxonomy of sound effects has been learned using unsupervised learning techniques.

At the first level, a hierarchical structure of the data was extracted and presented in Figure 4. Following from this, a decision tree was created and pruned, to allow for visualisation of the data, as in Figure 5. Finally a semantically relevant context was applied to data, to produce a meaningful taxonomy of sound effects which is presented in Figure 6. A semantic relationship between different sonic clusters was identified.

The hierarchical clusters of the data provide deeper understanding of the separating attributes of sound effects, and gives us an insight into relevant audio features for sound effect classification. We demonstrated the importance of the periodicity, dynamic range and spectral features for classification. It should be noted that although the entire classification was performed in an unsupervised manner, there is still a perceptual relevance to the results and there is a level of intuition provided by the decision tree and our semantic descriptors. Furthermore, the clustering and structure will be heavily reliant on the sound effects library used.

We also demonstrated that current sound effect classification and taxonomies may not be ideal for their purpose. They are both non-standard and often place sonically similar sounds in very different categories, potentially making it challenging for a sound designer to find an appropriate sound. We have proposed a direction for producing new sound effect taxonomies based purely on the sonic content of the samples, rather than source or context metadata.

In future work, validation of the results on larger sound effect datasets could aid in evaluation. By using the hierarchical clustering method, one can also produce a cophonetic distance between two samples. This would allow identification of how the distance can correlate with perceived similarity and may provide some interesting and insightful results. Further development of the evaluation and validation of the results, perhaps through perceptual listening tests, would be of beneficial to this field of research. It is also possible to look at the applications of hierarchical clustering towards other types of musical sounds, such as musical instrument classification. Hierarchical clustering is able to provide more information and context than many other unsupervised clus-

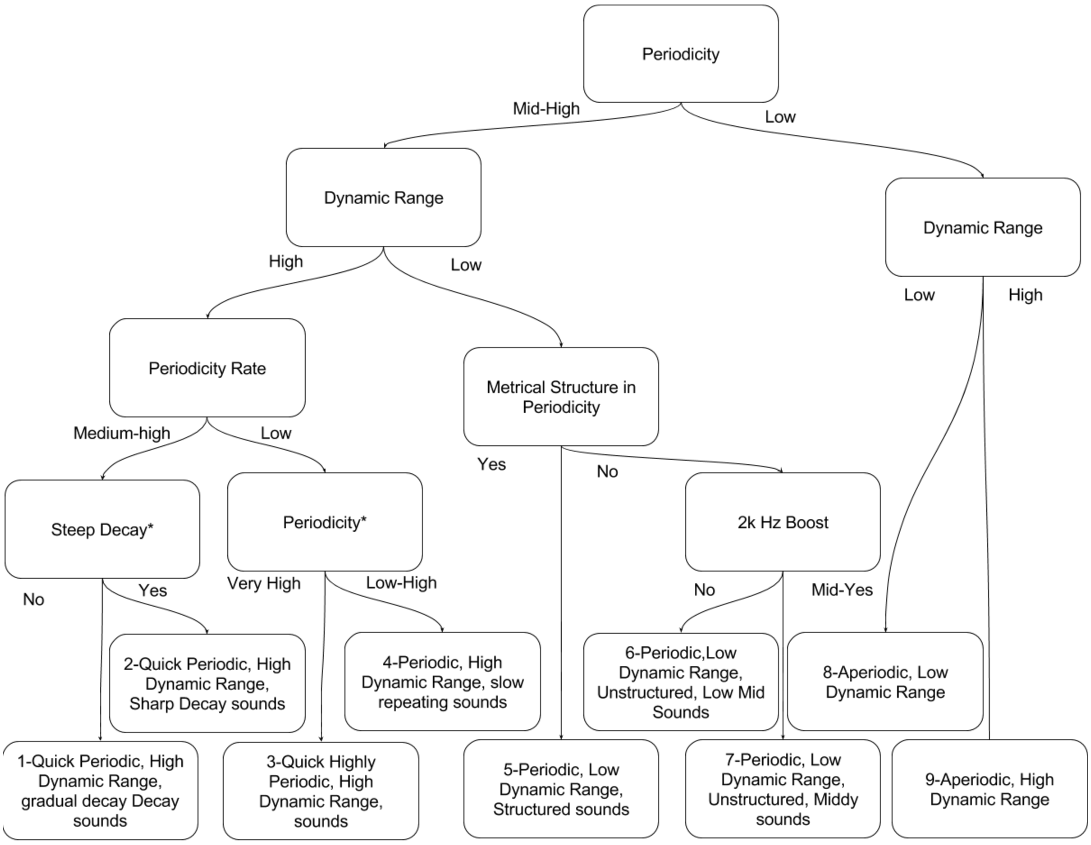
Periodicity
Mid-High
Low
Dynamic Range
Dynamic Range
High
Low
Low
High
Periodicity Rate
Metrical Structure in Periodicity
Medium-high
Low
Yes
No
Steep Decay*
Periodicity*
2k Hz Boost
No
Yes
Very High
Low-High
No
Mid-Yes
2-Quick Periodic, High Dynamic Range, Sharp Decay sounds
4-Periodic, High Dynamic Range, slow repeating sounds
6-Periodic,Low Dynamic Range, Unstructured, Low Mid Sounds
8-Aperiodic, Low Dynamic Range
1-Quick Periodic, High Dynamic Range, gradual decay Decay sounds
3-Quick Highly Periodic, High Dynamic Range, sounds
5-Periodic, Low Dynamic Range, Structured sounds
7-Periodic, Low Dynamic Range, Unstructured, Middy sounds
9-Aperiodic, High Dynamic Range

Figure 6: *Machine learned taxonomy, where each node separation point is determined by hierarchical clustering and text within each node is an semantic interpretation of the most contributing audio feature for classification. Each final cluster is given a cluster number and a brief semantic description. The ∗ represents a feature separation where the classification accuracy is less than 80%, never less than 75%.*

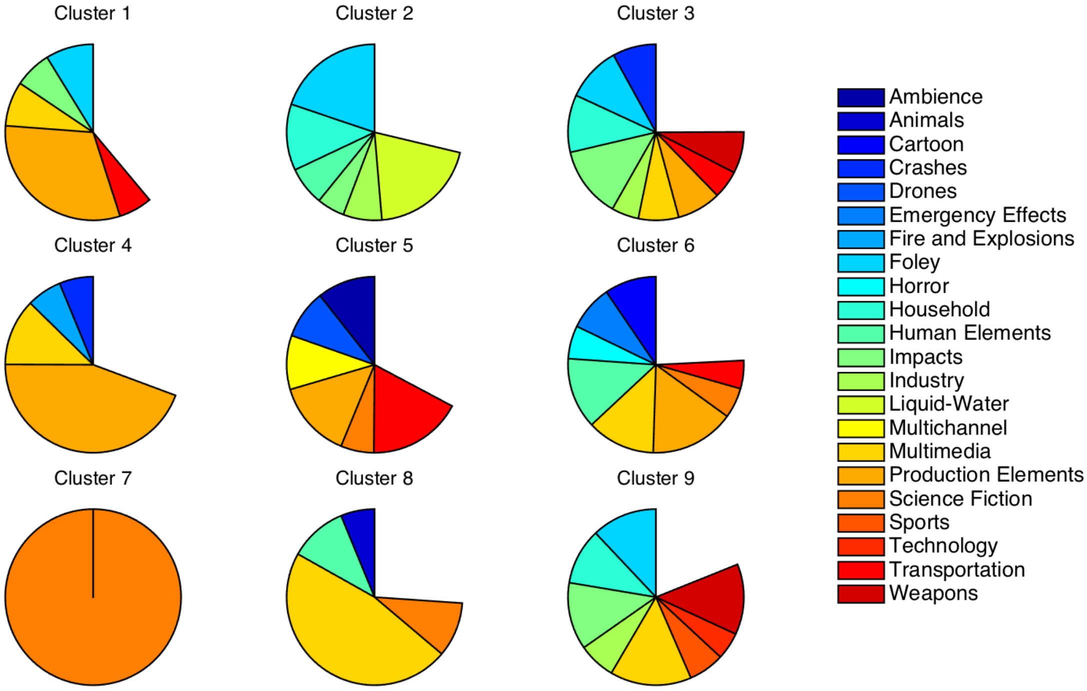

Figure 7: *Dataset labels per cluster, where all labels that make up more than 5% of the dataset were plotted*

tering methods. Further evaluation of clusters produced could be undertaken, as well as a deeper analysis into each of the identified clusters, to produce a deeper taxonomy.

7. ACKNOWLEDGEMENTS

This work was supported by EPSRC. Many thanks to anonymous internal reviews.

8. REFERENCES

[1] Erling Wold et al., "Content-based classification, search, and retrieval of audio," *IEEE multimedia*, vol. 3, no. 3, pp. 27–36, 1996.

[2] Grégoire Lafay, Nicolas Misdariis, Mathieu Lagrange, and Mathias Rossignol, "Semantic browsing of sound databases without keywords," *Journal of the Audio Engineering Society*, vol. 64, no. 9, pp. 628–635, 2016.

[3] R Murray Schafer, *The soundscape: Our sonic environment and the tuning of the world*, Inner Traditions/Bear & Co, 1993.

[4] Luigi Russolo and Francesco Balilla Pratella, *The art of noise:(futurist manifesto, 1913)*, Something Else Press, 1967.

[5] Luigi Russolo, "The art of noises: Futurist manifesto," *Audio culture: Readings in modern music*, pp. 10–14, 2004.

[6] Ian Stevenson, "Soundscape analysis for effective sound design in commercial environments," in *Sonic Environments Australasian Computer Music Conference*. Australasian Computer Music Association, 2016.

[7] PerMagnus Lindborg, "A taxonomy of sound sources in restaurants," *Applied Acoustics*, vol. 110, pp. 297–310, 2016.

[8] William W Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[9] Olivier Houix et al., "A lexical analysis of environmental sound categories.," *Journal of Experimental Psychology: Applied*, vol. 18, no. 1, pp. 52, 2012.

[10] James A Ballas, "Common factors in the identification of an assortment of brief everyday sounds.," *Journal of experimental psychology: human perception and performance*, vol. 19, no. 2, pp. 250, 1993.

[11] Brian Gygi, Gary R Kidd, and Charles S Watson, "Similarity and categorization of environmental sounds," *Perception & psychophysics*, vol. 69, no. 6, pp. 839–855, 2007.

[12] Kirsteen M Aldrich, Elizabeth J Hellier, and Judy Edworthy, "What determines auditory similarity? the effect of stimulus group and methodology," *The Quarterly Journal of Experimental Psychology*, vol. 62, no. 1, pp. 63–83, 2009.

[13] Monika Rychtáriková and Gerrit Vermeir, "Soundscape categorization on the basis of objective acoustical parameters," *Applied Acoustics*, vol. 74, no. 2, pp. 240–247, 2013.

[14] William J Davies et al., "Perception of soundscapes: An interdisciplinary approach," *Applied Acoustics*, vol. 74, no. 2, pp. 224–231, 2013.

[15] Iain McGregor et al., "Sound and soundscape classification: establishing key auditory dimensions and their relative importance," in *12th International Conference on Auditory Display*, London, UK, June 2006.

[16] Torben Holm Pedersen, *The Semantic Space of Sounds*, Delta, 2008.

[17] James Woodcock et al., "Categorization of broadcast audio objects in complex auditory scenes," *Journal of the Audio Engineering Society*, vol. 64, no. 6, pp. 380–394, 2016.

[18] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1041–1044.

[19] Davide Rocchesso and Federico Fontana, *The sounding object*, Mondo estremo, 2003.

[20] Edgar Hemery and Jean-Julien Aucouturier, "One hundred ways to process time, frequency, rate and scale in the central auditory system: a pattern-recognition meta-analysis," *Frontiers in computational neuroscience*, vol. 9, 2015.

[21] Justin Salamon and Juan Pablo Bello, "Unsupervised feature learning for urban sound classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 171–175.

[22] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 279–283, 2017.

[23] David J. M. Robinson and Malcolm J. Hawksfords, "Psychoacoustic models and non-linear human hearing," in *109th Audio Engineering Society Convention*, Los Angeles, CA, USA, September 2000.

[24] Dmitry Bogdanov et al., "Essentia: An audio analysis library for music information retrieval," in *ISMIR*, 2013, pp. 493–498.

[25] David Moffat, David Ronan, and Joshusa D. Reiss, "An evaluation of audio feature extraction toolboxes," in *Proc. 18th International Conference on Digital Audio Effects (DAFx-15)*, November 2015.

[26] David Ronan, David Moffat, Hatice Gunes, and Joshua D. Reiss, "Automatic subgrouping of multitrack audio," in *Proc. 18th International Conference on Digital Audio Effects (DAFx-15)*. DAFx-15, November 2015.

[27] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot, "Variable selection using random forests," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[29] Joe H Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[30] David Baehrens et al., "How to explain individual classification decisions," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1803–1831, 2010.

[31] Leo Breiman et al., *Classification and regression trees*, CRC press, 1984.