Maynooth University

Thesis

CS648

Semantic Analysis of Timbre Descriptors

Masters in Data Science

Student:

Ian Finnegan

18145400

Supervisor:

Dr. Joseph Timoney

Presented to the Maynooth University in Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Data Science and Analytics

Declaration

I hereby certify that the thesis I am submitting for assessment as part of the Master of Science in Data Science and Analytics degree is entirely my own original work, and has not been taken from the work of others, except where otherwise indicated. I am aware of the University's regulations concerning plagiarism, including those regulations concerning disciplinary actions that may result from plagiarism.

Signed: Ian Finnegan

Acknowledgements

I would first like to express my thanks to the Semantic Audio Lab in London and the Interactive Audio Lab at Northwestern University in Illinois for generously sharing, and giving me access to, their collected data, without which this report would not be possible.

I would further like to thank my supervisor, Dr.Joseph Timoney, to whom I owe a great gratitude to for his profound guidance, valued advice, and well appreciated critical analysis. This project and report would not have reached the level of quality that it currently stands at without his approachability, thoroughness, and kindness in regards to assisting others.

Finally, I would to thank my parents Annette and Alan, my sister Amy, and my family and friends, for their never-ending support, encouragement, and guidance. They are the shoulders I lean on, and their counsel and advice are integral to the person I am today and where I get to stand.

Thank you.

Abstract

Semantic Analysis of Timbre Descriptors

This report encompasses an attempt to quantify and parametrise qualitative perceptual descriptors used to describe the multidimensional characteristics of timbre changes by two crowdsourced groups consisting of audio experts and audio layman across four effects tools, equalisation, reverberation, compression, and distortion. The salient outcome being that timbre descriptor definitions are best represented in the form of dimensionality reduced principle components, that maintain over 70% of the variation of the data, constructed from frequency spectra extracted spectral features associated with an individual descriptor in each of the applied effects. This results in a method to distinguish between applied effects by spectral variance and spectral standard deviation and define descriptors in a comparative way using spectral feature loading correlations specific to a given descriptor with a specific applied effect.

Contents

List of Tables

List of Figures

1 Introduction

Timbre is the characteristic of sound that allows a listener to distinguish between tones, pitch, intensity, vocals, instruments, recording environments, and a range of other audio conditions. An auditory experience that has been studied, manipulated, and optimised to better the sounds produced as far back as music goes, but whose true nuance has only really been understood over the last hundred years. With that recent understanding, the reasons behind why timbre is so poorly understood has been suggested to be due to its multidimensional nature.

With no universal sensory vocabulary to define the characteristics of timbre, there is no standard units or parameters to quantify timbre in any reasonable or scientific way. Instead, timbre is described using a wide range of perceptual descriptors given by listeners whose qualitative nature isn't comparative, as one persons perception of a sound being "warm", isn't another's. However, this report will examine ways to optimise and parametrise these qualitative descriptors in way to quantify what a given descriptor means in relation to the sound, and associated spectral features, it is being used to describe.

To that end, data was collected from two groups across four applied effects. The two groups being that of, audio experts made up of sound engineers and recording producers, and audio layman made up of students and crowd-sourced participants willing to listen to and describes music. Both groups described music that had been affected by four effects tools, equalisation, reverberation, compression, and distortion, these tools are used professionally for the manipulation of timbre.

The equalisation tool allows a user to change the gain (amplitude) of each frequency band that makes up a sound. A reverberation tool can apply a small or large echo at a particular frequency changing the sound of it's environment. Compression reduces the overall dynamic range of a sound, changing the gain structure as a whole. And distortion overloads the gain structure to fill a spectrum with noise to produce a grunge type sound.

This reports is made up of four sections, the first being this introduction. Section 2. is background, in which there is an overview of the related research already completed in this area, and a description of the contribution made by this report and where it fits into the related works. Section 3. is the methodologies used to collect and clean the data used, as well as a description of the analysis techniques used and how they are applied. Section 4. is the results of all the analysis applied on the data and what that analysis determined. Finally, this reports ends with a Conclusion and a discussion about what potential future work can be done in this area both in terms of direct follow-ups and spin-off research.

2 Background

Related Works

The Acoustical Society of America's Terminology Standards Association (1960) defined timbre as a range of attributes consisting of a frequency spectrum, sound pressure and temporal characteristics that produce an "auditory sensation which enables a listener to judge that two nonidentical sounds, similarly presented" are in fact not the same sound. Thus, timbre is an incredibly important and fundamental part of the musical experience as it allows the listener to recognise the difference between a voice, a guitar, a piano, or a plethora of other vocals and instruments all equally tuned and vocalising or playing a note of the same fundamental pitch and volume.

Over the last 80 years a range of differing analysis has taken place in an attempt to produce a universal vocabulary that describes all of musical sounds. Unfortunately, due to the independent perceptual nature of people's creativity and taste it's very difficult to produce a set of descriptors or dimensions that represent timbre empirically. The prevailing way to describe timbre is via salient semantic dimensions usually produced with pairwise dissimilar attributes. A pairwise dissimilar dimension being a one-dimensional scale made up from a pair of descriptors, think of it like a thermostat, a one-dimensional scale that goes from descriptor cold up to descriptor hot. One of the earliest of which is Helmholtz (1954) whose study originally described timbre in terms of two pairwise dimensions pitch (low to high) and loudness (soft to loud).

This was followed up by having ”sonar technicians” on Solomon’s (1958) team grade recordings of sonar sounds on pairwise scales producing a seven-dimension solution. The most salient dimensions being that of magnitude (light to heavy, small to large), and clarity (clear to opaque, definite to indefinite). The rest being said to be aesthetic in nature (smooth to rough, dull to sharp). Gottfried von Bismarck (1974) undertook one of the first comprehensive analysis of timbre semantics using 30 verbal scales in order to rate 35 speech sounds. His study produced four orthogonal dimensions, volume (empty to full), texture (dull to sharp), colour (-less to -full) and diffusion (compact to spread). He further argued that since his texture dimension captured nearly half of the total variance of his data, it may be another distinguishing attribute between sounds similar to pitch and loudness. The study also found that musicians and engineers where more consistent than layman, however, he discarded descriptors based on arbitrary judgement of synonymy thus potentially forcing a more agreeable set of dimensions.

In Pratt and Doak (1976) 19 commonly used adjectives to describe synthetic tones were reduced to six by arbitrarily discarding synonyms and useless descriptors in terms of the authors opinion and then proposed a three-dimensional map composed of vision (dull to bright), temperature (cold to warm) and wealth (pure to rich). McAdams (1995) in his attempt at finding universal sound descriptors that map onto a set of pairwise scales noted that there are a variety of descriptors due to individual perception of sounds that are not mapped onto pairwise dimensions and thus two or three pairwise dimensions might not be adequate. This is further corroborated by Rossing (2002) who states that “loudness and pitch are a concept of lower dimensions”, allowing for ordering in terms of quite to loud via frequency alone but that timbre is a much more complex sound property and would require greater multidimensional mapping and dimensions that go beyond pairwise scaling. This can also be seen in a more recent study by Saitis (2017) who states, “individual descriptors should be thought of as basic elements of semantic knowledge” and thus are not ”fully meaningful on their own” and only meaningfully informative when placed into more extensive semantic dimensions. Thus, the diverse timbre descriptor vocabulary may be due to many seemingly dissimilar words actually sharing the same perceptual dimension

Moravec and Stepanek (2003) was one of the first big studies of timbre in the 21st century compiling a crowdsourced list of words typically used to describe the timbre by conductors, composers, engineers, teachers, and bowed-string, wind, and keyboard musicians. The most frequently used descriptors were sharp, gloomy, soft and clear, with bowed-string players preferring to use sweet and warm more frequently, while wind performers were more often associated with the descriptor narrow. The 30 most frequent descriptors were used to produce a four dimensional map consisting of a more complex relationship than a simple pairwise scale. These dimensions consisted of vision (bright/clear/cloudy/dark), texture (hard/sharp/dull/delicate/soft), volume (wide/narrow/close/far/empty), and temperature (hot/cold/warm). Stepanek (2006) followed up with a similar study that also determined the dimensions vision, texture, and volume but changed the last one to hearing (noisy/rustle/ringing). He further suggested that sounds described as sharp are just sounds that are bright and rough. Though, other studies into timbre semantics have shown that bright, rough, and sharp consistently show up in plethora of different dimensions independently.

Over the last 15 years there has been a sea of studies into timbre, Mecklenburg (2006) attempted to examine more complicated descriptor formats by mixing descriptors with applied effects such as “warm-hi-cut” but the labels were chosen by the author and musicians considered the terms too technical. Disley (2006) then tried to pass the complexity of timbre description onto the dimensions themselves by widening what could be defined in each dimension. Using fifteen parameters to produce a four-dimensional semantic space by studying notes played by twelve orchestral instruments of common pitch, Disley determined dimensions of aesthetic (bright/thin/harsh/-clear/dull), purity (percussive/electric/live), material (wooden/metallic) and evolving (changing in time). Sarkar (2007) investigated a general taxonomies of sound descriptors. Sundaram (2007) focused on onomatopoeic descriptors, rather than the broader range of all possible descriptors however it was a small biased study of his four lab members.

Sabin (2009) chose four labels (“bright”, “warm”, “tinny”, “dark”) in an attempt to produce an interface that lets users control and explore different equalisations. Howard (2009) independently reproduced Disley’s result of aesthetic, purity, material, and evolving. Sound engineer David Huber (2010) went beyond just the frequency-based equalisation studies that all the previous investigations have been to describe effects produced by recording and production equipment and analyse compression and distortion properties. Ferrer (2011) examined what descriptors are used to describe songs played on the radio and what clusters are produced by those descriptors. Zacharakis (2012) used multidimensional scaling to study 30 parameters of 23 different ”notes from acoustic, electric, and electronic instruments” and produced a three-dimensional solution containing texture (soft/rounded/warm/rough/harsh), mass (dense/rich/full/thick/light), and luminance (brilliant/sharp/deep).

At this point a plethora of studies have been done with regards to audio experts, lab members and research authors but very few had been done with regards to layman. This changed with Cartwright and Pardo (2013, 2014), and Seetharaman and Pardo (2014, 2016) who crowdsourced descriptors from layman who would listen to pieces of music and describe it. This was done with music pieces that were using equalisation, compression, reverberation, and distortion effects. However, they only ever analysed tallied counts of occurrences to determine popularity of certain descriptors for certain effects. The most recent major timbre study was done by Wallmark (2018) who undertook a corpus linguistic analysis of the descriptor vocabulary used in orchestra books, manuals, and music sheets. However, he also only used tallied occurrences, determining seven categories that consisted of crossmodal correspondence (borrowed from other senses), affect (emotion and aesthetics), action (physicality, movement), acoustics, matter (weight, size, shape), onomatopoeia, and mimesis (sonic resemblance). This in turn produced three dimensions, activity (action and mimesis), sensory (crossmodal and acoustics), and material (onomatopoeia and matter).

Contribution

As seen above finding universally defined sound descriptors that map onto a set of dynamic dimensions in attempts to empirically parameterise and define timbre is an incredibly difficult but sought-after solution. However, many problems and loose ends still remain, this investigation hopes to make attempts at remedying a few of them. The first of which is the lack of limited bias crowdsourced data being used. A great deal of the studies outlined above have a tendency to involve author chosen words, author associated lab members, and author discarded descriptors which all force a solution which won't truly represent the timbre it's trying to explain. A secondary problem is that none of the previous studies directly compared layman perception of sounds with that of experts in the associated fields. The closest studies got to this was von Bismarck (1974) who noted that experts were more consistent with their fellow experts. This is further demonstrated by Toulson (2003) who asserts that "timbre descriptor terms are not widely understood by either musicians or the general public". These problems will be solved by only using crowdsourced data collected from a group of experts and a group of layman, whose parameters that make up common descriptors will be directly compared and examined for similarities or lack thereof. More on the data itself will be outlined below in the Data section of Methodologies.

A third problem exists in the form of only examining the popularity of descriptors via tallied counts, and the associated distributions and usage clusters determined by those counts. This will be solved by widening the analysis done on the collected data to also include comparisons of the parameter space that is associated with the common descriptors, examining the agreement scores between expert and layman common descriptors and the extraction and comparison of spectral feature data and where common descriptors fit in feature space via dimensionality reduction. All of these analysis techniques are also outlined in greater detail below. The final problem this investigation will examine is that none of the studies above compare timbre descriptors across timbre effects such as equalisation, reverberation, compression, and distortion. Previous studies only ever examine them independent of each other, this will be solved by comparing the contributions of the extracted spectral features on each of the timbre effects and examining where the associated common descriptors fit in this cross-effect space. Further problems and research possibilities also exist and will be outlined later in the Future Work section.

3 Methodology

3.1 Data

Collection

The raw expert and layman data was collected from two data banks. The first, the expert data, was collected from the Semantic Audio Labs (SAL) in London which is a user group for sound engineering and media production data and tools. From SAL the Semantic Audio Feature Extraction (SAFE) plugin database was acquired, which is composed of datasets for the four effects; equalisation, reverberation, compression, and distortion. The SAFE plugin is part of a studios Digital Audio Workstation (DAW), which is a modern digital sound desk used in most studios for recording, editing, mastering, and overall production of music and audio projects. As an engineer or producer edits a piece of music, via the effects tools, the timbre is changed. The SAFE plugin then records the changes made and asks the user to enter a descriptor that describes the sound the user is intending to produce.

The second, the layman data, was acquired from the Interactive Audio Lab (IAL) at the Computer Science Department of Northwestern University in Illinois. The data was collected by utilising Amazon's Mechanical Turk, a crowdsourcing website to perform discrete tasks. Regular people unassociated with the sound production industry would listen to a piece of music, that piece would then have a series of different effects applied to it and the person could then swap back and forth between the pre-effect and post-effect versions. They would then be asked to describe, in there own words, how the piece of music sounded. This was done for all four timbre effects and a range of applications.

Cleaning

	Expert		Layman	
Effect	Before	After	Before	After
Equalisaion	1700	1386	1596	918
Reverberation	441	230	682	449
Compression	468	468	460	383
Distortion	309	309	604	398

Table 1: Number of Entries per Effect Before and After Cleaning

Due to the fact that both the SAFE and IAL datasets allow experts and layman, respectively, to describe timbral transformations in terms of natural language, in many cases there were misspellings, an assortment of random letters or numbers, a description of why the changes were made, a label for the sound, or a range of suffixes. These problems were fixed by removing entries whose descriptors were illegible, unusable, or obscure. The remaining entries were fixed in accordance with a Porter Stemmer (Porter, 1980), which groups terms with shared meanings by applying stemming conditions on the suffixes. This combines descriptors like *warm*, *warmer*, and *warmth*, along with misspellings like *warnth* into the single descriptor *warm*. As can be seen in the table above the number of entries in the data that get used in the analyses is different from the number of entries in the raw data once these cleaning conditions are applied.

Equalisaion	Reverberation	Compression	Distortion
Descriptor	Descriptor	Descriptor	Descriptor
Frequency	Dampening Freq	Threshold Freq	Frequency
Gain	Density	Gain	Gain

Table 2: Relevant Column Variables per Effect

Once the entries of the datasets were cleaned the column variables were then also trimmed down to only the most relevant to the analyses. As seen in the table, that basically consists of a descriptor column, a frequency column, and gain column. The damping and threshold frequency is the equivalent variable to frequency for the those effects. Similarly, density is the equivalent variable for the gain for it's associated effect. Other variables were also present in the raw data such as decay and predelay for reverberation, attack and ratio for compression, and knee and bias for distortion. However, those variables where not recorded in both the expert and layman datasets, therefore, couldn't be effectively compared.

3.2 Analysis Techniques

Spectra and Spectral Features

A spectrum is the distribution of a given spectral value in a specific domain. In the case of audio, it's the distribution of gain amplitudes of each frequency component in frequency space, this shows the amount of gain within each frequency band over a range of frequencies. The gain is the change in power of each signal at each frequency band from the input signal to the output signal by adding voltage to said signal. When a raw sound comes in an engineer can add or remove voltage at a specific frequency and therefore, increase or decrease the gain at that frequency and thus, change the timbre of the sound. To analyse this the overall spectrum can be examined for common distributions across descriptors, or its spectral characteristics, called spectral features, can be utilised.

Spectral Feature	Feature Equation
Centroid	$\mu_1 = \frac{\sum_{k=b_1}^{b_2} f_k s_k}{\sum_{k=b_1}^{b_2} s_k}$
Variance	$\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1) s_k}{\sum_{k=b_1}^{b_2} s_k}$
Standard Deviation	$\mu_2 = \sqrt{\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1) s_k}{\sum_{k=b_1}^{b_2} s_k}}$
Skewness	$\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^3 s_k}{(\mu_2)^3 \sum_{k=b_1}^{b_2} s_k}$
Kurtosis	$\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^4 s_k}{(\mu_2)^4 \sum_{k=b_1}^{b_2} s_k}$
Roll Off	$\sum_{k=b_1}^{i} \lvert s_k \rvert = K \sum_{k=b_1}^{b_2} s_k$
Flatness	$\frac{(\prod_{k=b_1}^{b_2} s_k)^{\frac{1}{b_2 - b_1}}}{\frac{1}{b_2 - b_1} \sum_{k=b_1}^{b_2} s_k}$
Crest	$\frac{\max(s_k)}{\frac{1}{b_2 - b_1} \sum_{k=b_1}^{b_2} s_k}$
Slope	$\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_f)(s_k - \mu_s)}{\sum_{k=b_1}^{b_2} (f_k - \mu_f)^2}$

Table 3: Spectral Features and Associated Equations. Where f_k is the frequency at bin k, s_k is the spectral value at bin k, b_1, b_2 are the bin edges, K is the 95% energy threshold, μ_f is the mean frequency and μ_s is the mean spectral value.[30][38]

In the above table the spectral features and how to calculate them can be seen, where each feature numerically represents a different characteristic aspect of a spectrum. The centroid is a frequency weighted sum that represents the overall centre of a distribution. The variance and standard deviation both measure the spread of a spectrum with the former representing the average spread of a spectrum and the latter representing the instantaneous bandwidth around a given centroid. The skewness measures the symmetry of a spectrum around a centroid, and the kurtosis indicates the peakiness of a spectrum representing its flatness or Gaussianity. The roll off is the point on a spectrum at which 95% of the total energy exist and flatness is the ratio of the geometric mean to the arithmetic mean and is used as an indication of noise in a signal. The spectral crest is another indication of noise using the ratio of the maximum gain of a spectrum to the arithmetic mean, and finally the slope represents the decrease across a spectrum when there's more energy in the lower frequencies than the higher ones.

Hierarchical Clustering

Hierarchical clustering is a statistical method that attempts to create non-overlapping subgroups of unknown membership with the points from a given dataset by some similarity criterion, in this case, the effects datasets with groups of semantically related descriptors. It does this by treating each data point as a singleton cluster, and then successively merging until all points have been merged, iterating towards a solution by taking the best step at each stage to maximise in group homogeneity and between group heterogeneity. Another clustering method called k-means clustering is also a viable option, however k-means requires an initial pre-specified number of clusters and also hierarchical clustering has better graphical presentation in the form of dendrograms. Clustering algorithms require two things, a distance matrix, usually either euclidean or manhattan, and a linkage method, usually one of single, complete, average or ward.

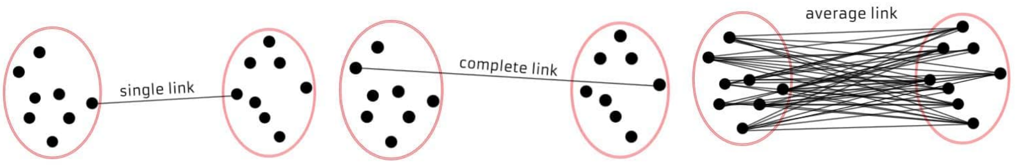
Distance	Equation
Euclidean	$\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$
Manhattan	$\lvert x_1 - x_j \rvert + \lvert y_i - y_j \rvert$

(a) Distance Methods[32]

Linkage	Equation
Single	$\text{Min}_{x \in C_1, y \in C_2} d(x, y)$
Complete	$\text{Max}_{x \in C_1, y \in C_2} d(x, y)$
Average	$\frac{1}{\lvert x \rvert \lvert y \rvert} \Sigma_{x \in C_1} \Sigma_{y \in C_2} d(x, y)$
Ward	$\Sigma_{x \in C_1} \Sigma_{y \in C_2} \lvert\lvert x - y \rvert\rvert^2$

(b) Linkage Methods[32]

Table 4: Equations for Clustering Methods. Where x and y are points in the data, and C_1 and C_2 are clusters of the data



(a) Single (b) Complete (c) Average

Figure 1: Types of Clustering Linkage Methods[32]

A distance matrix is a table showing the distances between pairs of points calculated via euclidean or manhattan method (Stahl et al, 2011). The euclidean distance is simply the straight line distance between two points, whereas the manhattan distance is the distance by taking right angles between points. In more recognisable terms, the former is as the crow flies, and the latter is like following the street pathways. Once the distance between points is determined, how to link one cluster to another is the next part, this is done via a linkage method.

Single linkage links two clusters by the smallest distance between points of those cluster, however it fails when dealing with poorly separated clusters. Complete linkage links clusters by the largest distance between points of the clusters. Average linkage links by the average of the distance between the points of one cluster and the points of another. Finally, the ward linkage links clusters by which combinations produce the lowest variances. Iterating through all possible combinations allows for a determination of the clustering structure, the closer to one the structure is the tighter the clusters are going to be, thus, the greater the homogeneity within clusters.

To choose the most appropriate linkage method, it is useful to compare their respective pros and cons. While single linkage is very efficient at producing clusters of different sizes and shapes, it is also very sensitive to noise in the data. This means that it fails when the data, and associated groupings, are poorly separated. Average linkage is slower in computation than single linkage, but it produces a more robust solution since it isn't as sensitive to noise. However, it is biased towards global patterns between clusters over local similarities within clusters. Similarly, complete linkage has the same bias, and produces an equally robust solution as average linkage, but is even more computationally taxing. The ward method is somewhat a happy medium which produces a global solution, but due to similar variance, also produces tightly bound spherical clusters that are less sensitive to both noise and outliers. It will be seen later that the ward linkage method also produces the strongest clustering structure for the data.

Discriminant Analysis

Discriminant analysis is another statistical method that differs from clustering but is also used to create subgroups of non-overlapping points. While clustering is an unsupervised method with no predefined conditions, discriminant analysis is a supervised method that classifies points into predetermined classes using scores determined by associated characteristic column variables. There are three main classification models, Logistic Regression, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA), all three of which use probability distributions to sort points into classes.

Method	Equation
Logistic	$\frac{\exp^{\hat{\beta_0}+\hat{\beta_1}X}}{1+\exp^{\hat{\beta_0}+\hat{\beta_1}X}}$
LDA	$x^T\sum^{-1}\hat{\mu}_k-\frac{1}{2}\hat{\mu}_k^T\sum^{-1}-\hat{\mu}_k+log(\hat{\pi}_k)$
QDA	$-\frac{1}{2}x^T\sum_k^{-1}x+x^T\sum_k^{-1}\hat{\mu}_k-\frac{1}{2}\hat{\mu}_k^T\sum_k^{-1}\hat{\mu}_k-\frac{1}{2}log\lvert\sum_k\rvert+log(\hat{\pi}_k)$

Table 5: Equations for Classification Methods. Where β_0 and β_1 are model parameters, X is a column variable, k is a given class, μ_k is the mean of a variable for class k, $\sum$ is the covariance of the classes, whereas $\sum_k$ is the covariance of the variable for class k, and π_k is the prior probability that x belongs to class k.[36]

Logistic regression uses maximum likelihood, which determines model parameters from the data to give one class a probability of one and the other a probability of zero, so when an individual data entry enters it's variable's value into the model the probability result can place it into one class or the other. LDA determines its scores by obtaining a linear combinations of the variables, so when a point enters the model it is then placed into the class that has the highest associated score. QDA computes scores in a similar way to LDA, however, the model equation has more terms to account for each variables individual variances since LDA assume a common variance across the classes.

To choose the best method it's necessary to look at what potential disadvantages may hinder their application when compared with the other methods. Such as, when classes are well-separated the logistic regression model becomes unstable, in these cases both LDA and QDA perform better. LDA and QDA further perform better when the number of entries in a dataset are smaller and are also preferred over logistic regression when there are more than two response classes. A drawback of both LDA and logistic regression is that they assume equality of variance in the classes, therefore produce straight line boundaries between classes, this assumption is relaxed in QDA allowing for curved boundaries for a more dynamic distribution of classes.

Agreement Scores

Due to the peoples individual perception of sound and the semantics they use to describe that sound, one persons *warm* is not another persons *warm*. It is therefore necessary to determine how varied given descriptors are from one person to another, and thus, which descriptors have the most widely-agreed-upon meanings. To do this two method were used, Frechet Distance for continuous data and the normalised covariance for discrete data.

Method	Equation
Frechet Distance	$\mathrm{Min}(\mathrm{Max}(d(P(x_i), Q(x_j))))$
Normalised Covariance	$\left\lvert \dfrac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}} \right\rvert$

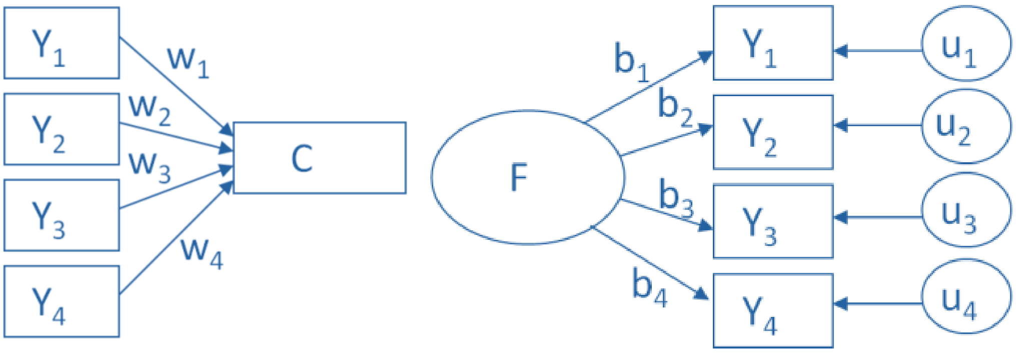
Table 6: Equations for Agreement Score Methods. Where $P(x_i)$ is the point x_i on curve P, $Q(x_j)$ is the point x_j on curve Q, rg_X and rg_Y are rank column variables, and σ_{rg_X} and σ_{rg_Y} are the standard deviations of those variables.[23]

Frechet distance is most commonly described as the minimum length of leash connecting a person to their dog where both are technically walking different paths but are in fact walking, overall, the same path. It is understood as the minimum maximum distance of two points on two paths. It can be calculated by determining the maximum distance between each point of one curve and every point on anther curve, then the minimum distance of all the maximum distances is the frechet distance. The lower the frechet distance the more similar the paths are and the higher the frechet distance the more dissimilar the paths are.

The normalised covariance, which can also be considered as the correlation between the two column variables, is a measure of how related the variables are to each other. While the most popular correlation method is Pearson's (Sayago at al, 2006), since the data is both non-continuous and non-linear, it is more appropriate to use Spearman's rank as it makes no assumptions about the distribution of the data. With correlation the closer a descriptor's value is to ±1 the more agreement there is about that descriptor, whereas the closer the value is to zero, the lower the agreement is about that descriptor.

Dimensionality Reduction

There are two main dimensionality reduction techniques which produce similar results and only differ in their core ideology, they are Principle Component Analysis(PCA) and Factor Analysis(FA) (Usman et al, 2017). Both methods reduce the number of variables in a manner that maximises the variability being represented by fewer dimensions and are indifferent to multicollinearity effects. Applying these methods allows for the shifting of space to maintain a high level of the data's information and do greater analysis in lower dimensionality when dealing with datasets that has many column variables.



(a) Principle Componant Analysis

(b) Factor Analysis

Figure 2: Dimensionality Reduction Methods

The salient difference between PCA and FA is that FA assumes the existence of factors underlying the observed data whereas PCA identifies dimensions that are combinations of the original variables. In PCA, the resulting components are orthogonal linear combinations of the original variables that maximise the total variance of the data in lower dimensionality. On the other hand, in FA, the factors are linear combinations that only maximise the shared variance of the data as it assumes this will determine the underlying dimension that the original variables represented.

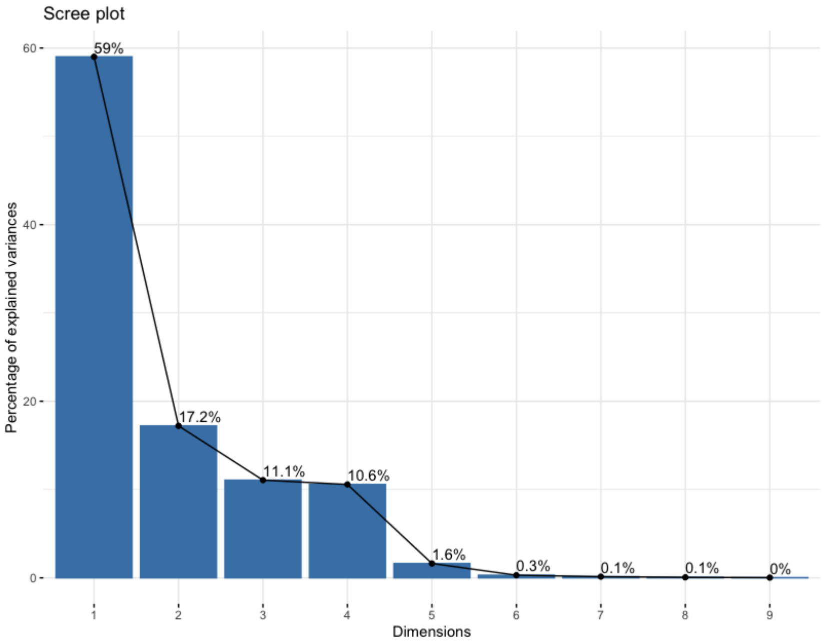


Figure 3: Screeplot of Equalisation PCA

While they both can be used interchangeably, results are usually more stable when FA is used for identifying the underlying factors that caused the original variables, and PCA is used to reduce correlated observed variables to a smaller set of independent dimensions. Since PCA will be used for the following analysis it's important to choose the appropriate minimum number of principal component that is enough to maintain a reasonable amount of the variation in the data. It's difficult to decide which principal components are enough but can be made easier via a screeplot, a plot of the eigenvalues (percentage of variation) per component. As seen in the plot above, components one and two cumulatively take into account most of the original data's information and can be use for most of the subsequent analysis, with more detailed analysis at most requiring the first four components, after which the sharp drop off in variation shows that little is accounted for in the later components.

4 Results

4.1 Equalisation (EQ)

Descriptor Occurrence

Expert			Layman		
Descriptor	n	Decimal	Descriptor	n	Decimal
warm	545	0.39	warm	65	0.07
bright	537	0.39	cold	34	0.04
bass	22	0.02	soft	29	0.03
low	18	0.01	loud	26	0.03
airy	16	0.01	happy	22	0.02
thin	16	0.01	bright	20	0.02
clean	14	0.01	clean	17	0.02
crisp	10	0.01	soothing	17	0.02
full	9	0.01	harsh	16	0.02
boxy	8	0.01	heavy	15	0.02
thick	8	0.01	cool	14	0.01
deep	7	0.01	smooth	14	0.01
muddy	7	0.01	calm	13	0.01
punchy	7	0.01	hard	11	0.01
bite	6	0.00	beautiful	10	0.01

Table 7: Top 15 Occurring Timbre Descriptors Used by Expert and Layman for Describing Equalisation

After cleaning the equalisation data, seen previously in Table 1., the remaining entries of 1386 and 918, for expert and layman respectively, can be broken down into the most occurring descriptors used. The top 15 of which can be seen in Table 7, representing 89% of the descriptors used by experts and 35% of the descriptors used by layman. This shows that when it comes to how equalisation effects are described, experts are much more concise, describing the bulk of the effect with fewer descriptors, whereas laymen have a much more diverse range of descriptors.

There is also a difference in what the descriptors are being used for, with the most frequent expert descriptors relating more to the attributes of the effect over how those attributes make the listener feel, as seen where layman use descriptors such as ”happy”, ”soothing”, and ”beautiful”.

Given this discrepancy between what experts and layman choose to describe, it should be examined what both groups actually mean when they use specific terms, and whether the meanings of shared terms represent similar attributes. To that end, Table 8. below is a list of the shared descriptors used by both groups.

		Expert		Layman	
Descriptor		n	Decimal	n	Decimal
1	warm	545	0.39	65	0.07
2	bright	537	0.39	20	0.02
3	bass	22	0.02	1	0.00
4	low	18	0.01	2	0.00
5	airy	16	0.01	3	0.00
6	clean	14	0.01	17	0.02
7	crisp	10	0.01	8	0.01
8	full	9	0.01	2	0.00
9	boxy	8	0.01	1	0.00
10	thick	8	0.01	1	0.00
11	deep	7	0.01	6	0.01
12	muddy	7	0.01	9	0.01
13	punchy	7	0.01	3	0.00
14	bite	6	0.00	1	0.00
15	aggressive	5	0.00	4	0.00
16	tight	5	0.00	1	0.00
17	fat	4	0.00	2	0.00
18	harsh	4	0.00	16	0.02
19	nice	4	0.00	3	0.00
20	dark	3	0.00	8	0.01
21	sharp	3	0.00	8	0.01
22	damped	2	0.00	1	0.00
23	dull	2	0.00	3	0.00
24	flat	2	0.00	3	0.00
25	quiet	2	0.00	5	0.01
26	smooth	2	0.00	14	0.01
27	soft	2	0.00	29	0.03
28	sweet	2	0.00	5	0.01
29	tender	2	0.00	1	0.00
30	big	1	0.00	1	0.00
31	fresh	1	0.00	2	0.00
32	good	1	0.00	3	0.00
33	light	1	0.00	6	0.01
34	lively	1	0.00	1	0.00
35	loud	1	0.00	26	0.03
36	pure	1	0.00	1	0.00
37	strong	1	0.00	2	0.00
38	twangy	1	0.00	1	0.00
39	vibrant	1	0.00	1	0.00

Table 8: Expert and Layman Common Descriptors for Equalisation

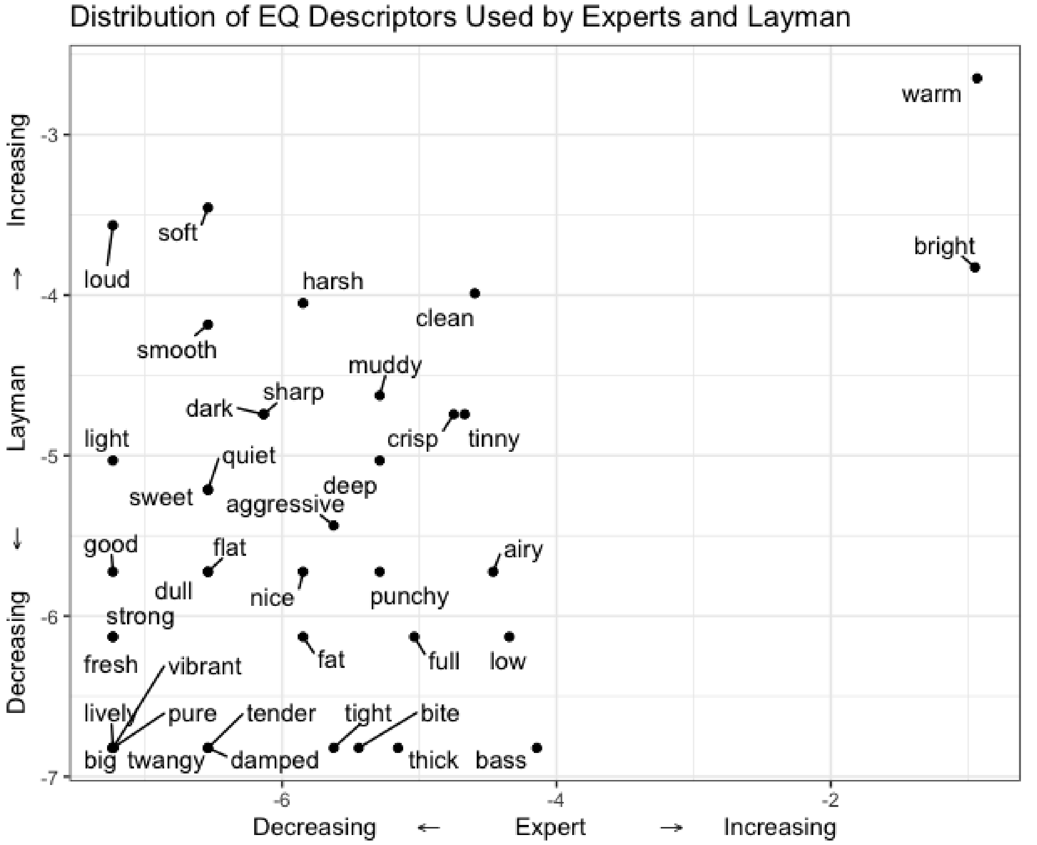


Figure 4: Distribution of Descriptors Used by Experts and Layman for Equalisation

The distribution of these common descriptors is shown in Figure 4. above, with the frequency of occurrence for expert and layman on the x and y axis respectively. Those with the most usage by both groups can be seen in the top right corner such ”warm” and ”bright”, with those least used shown in the bottom left corner such ”vibrant” and ”twangy”. It can also be seen that some descriptors are associated more strongly with one group than the other, such as ”loud” and ”soft” for layman and ”thick” and ”bass” for expert. Examining how these common descriptors bunch together based on their usage can further aid in determining a sort of synonymy or pairwise application of usage by their respective groups.

To start, which method to cluster by is established using the agglomerative coefficient to determine the strongest clustering structure by measuring the dissimilarity of the first cluster divided by the dissimilarity of the final cluster. The results shown in Table 9. for both expert and layman data, indicate the Ward linkage method will produce the tightest clusters with the most similarities.

	Average	Single	Complete	Ward
Expert	0.998	0.998	0.998	0.999
Layman	0.967	0.963	0.966	0.968

Table 9: Strength of Cluster Structure by Clustering Method

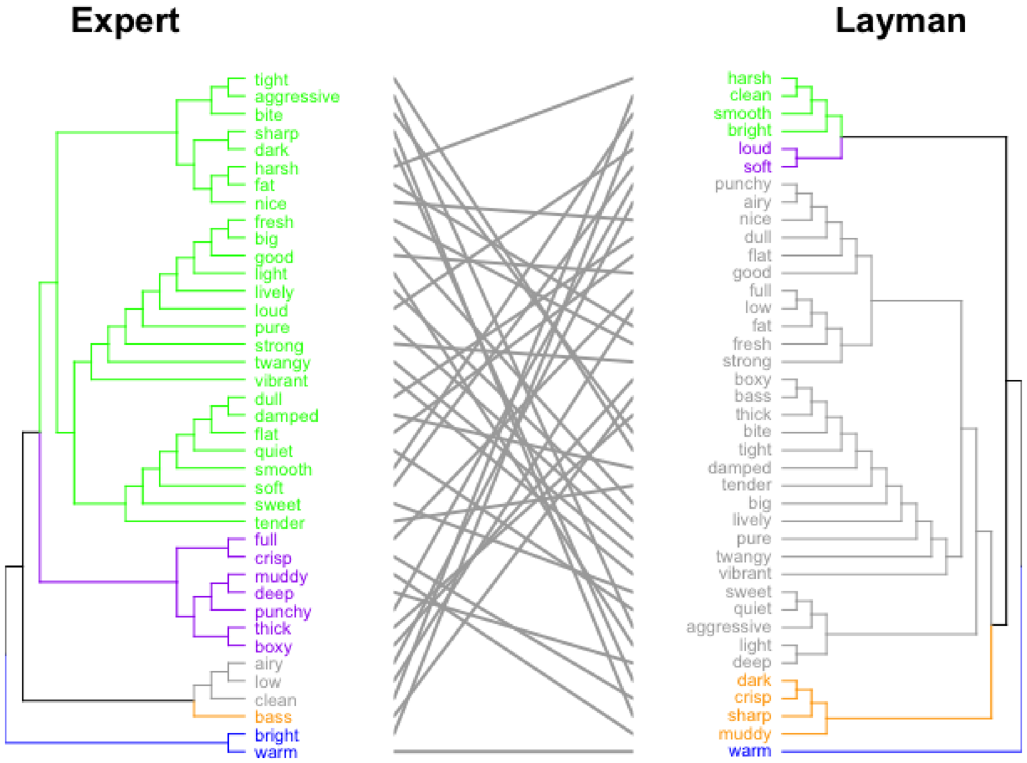


Figure 5: Clustering of Equalisation Descriptors Based on Expert and Layman Usage

It can be seen in Figure 5. that there are some usage clustering in terms of synonymy for the expert group with pairs such as "warm" and "bright", "thick" and "boxy", and "dull" and "damped". Whereas layman clusters produce a more dissimilar pairwise relationship with pairs such as "full" and "low", "loud" and "soft", and "harsh" and "clean". However a more in depth analysis is required to understand what these descriptors mean and how they relate to each other.

Descriptor Parameter Space

To begin a deeper analysis of the common descriptors the place to start is simple parameter space. In Figure 6. through to Figure 11. a selection from the top 10 occurring shared descriptors spectra can be seen for both experts and layman, along with a QDA model that predicts which group dominates in a given area of the frequency space.

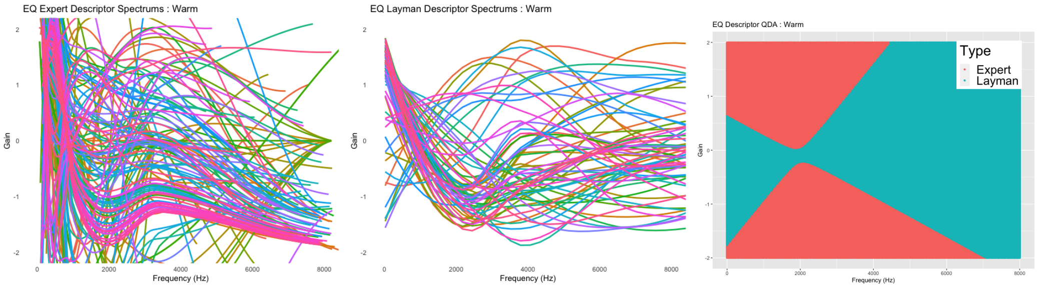
In Figure 6. (a) and (b) the expert and layman, respectively, spectra for descriptor Warm, an overall pattern can be seen where there is a +2 gain rise pre 1000Hz, a -2 gain drop around 2000Hz, followed by another rise to -1 gain at 4000Hz, and then a slow gain drop to -2 over 4000Hz to 8000Hz for experts. A similar pattern can be seen in the layman spectra with a nearly +2 gain rise pre 1000Hz, a -1.5 gain drop around 2000Hz, a rise to -1 gain at 4000Hz, but then a slow gain rise over 4000Hz to 8000Hz.

The major differences between the two sets of spectra are that the experts place their gains at more extreme levels than the layman, and that the layman try to lift their spectra back up towards the end whereas the experts just leave each parametric frequency alone only changing the ones needing change.

The QDA model produced by combining both the expert(in red) and layman(in blue) "warm" datasets, seen in Figure 6. (c), predicts the frequency space domination. The points coloured red had more expert points at it or closer to it than layman points, and the points coloured blue had more layman points at it or around it than expert points. What this does, is show where in frequency space expert definitions of descriptors dominate versus where layman descriptor definitions dominate.

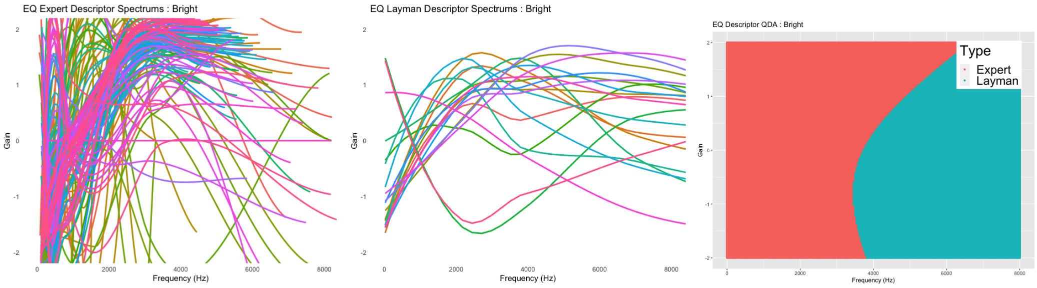
The gain structure breakdown discussed previously comparing Figure 6. (a) to (b) is reflected together in the QDA model Figure 6. (c). It shows the experts occupying the extremes around the 1000Hz and 2000Hz frequencies, and then steadily dropping off as it increases towards higher frequencies occupying less of the frequency spectrum. It further shows that layman spectra dominate space with lesser gain extremes occupying the middle gains. Only increasing in frequency space occupation as the frequencies increase into the higher 4000Hz to 8000Hz range.

Similarly, gain structure breakdowns and frequency space domination can be made for the other descriptors. Therefore, specific descriptor definitions for expert and layman common descriptors can be made and compared for Figure 7. through Figure 11.



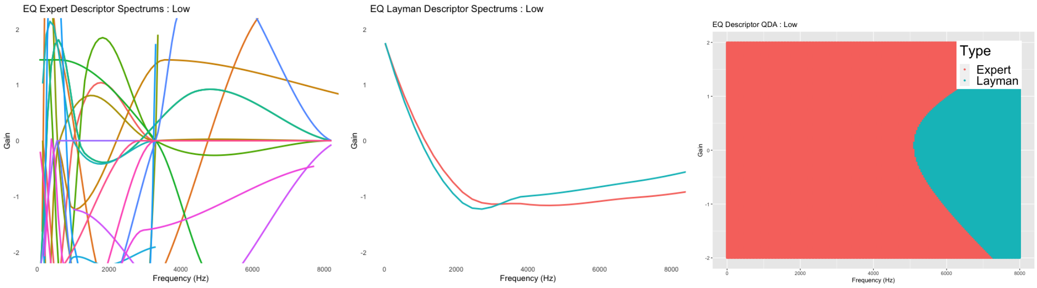
(a) Expert Warm Spectrum (b) Layman Warm Spectrum (c) Warm QDA

Figure 6: Warm Spectrums and QDA for Experts and Laymans



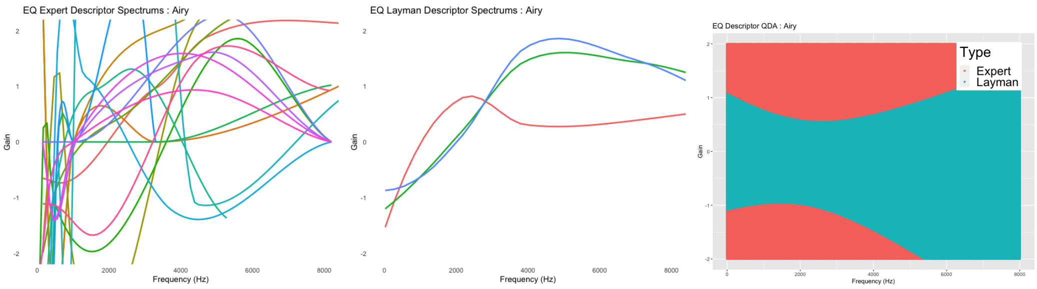
(a) Expert Bright Spectrum (b) Layman Bright Spectrum (c) Bright QDA

Figure 7: Bright Spectrums and QDA for Experts and Laymans



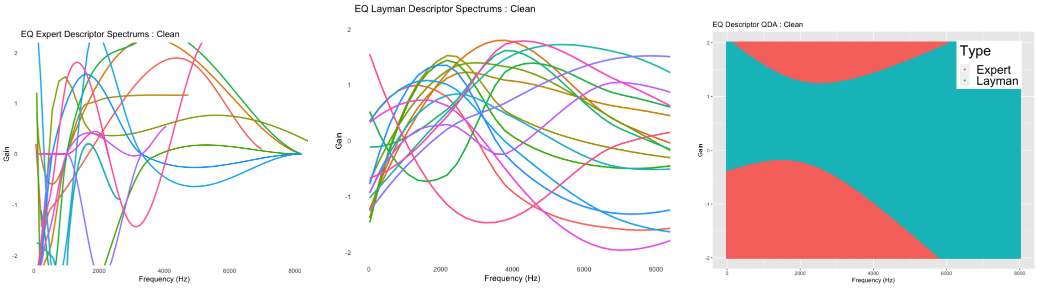
(a) Expert Low Spectrum (b) Layman Low Spectrum (c) Low QDA

Figure 8: Low Spectrums and QDA for Experts and Laymans



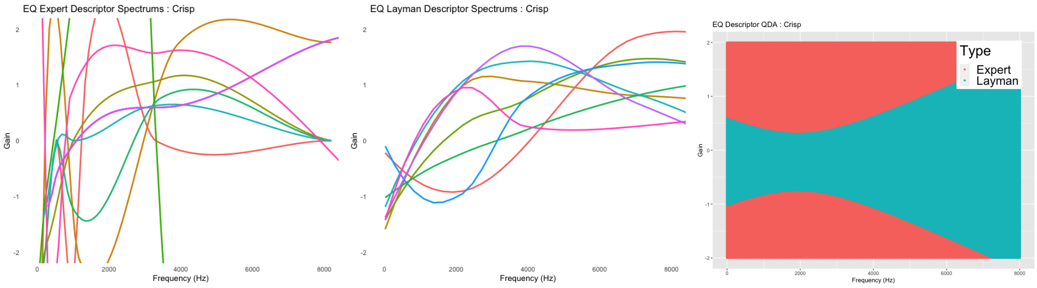
(a) Expert Airy Spectrum (b) Layman Airy Spectrum (c) Airy QDA

Figure 9: Airy Spectrums and QDA for Experts and Laymans



(a) Expert Clean Spectrum (b) Layman Clean Spectrum (c) Clean QDA

Figure 10: Clean Spectrums and QDA for Experts and Laymans



(a) Expert Crisp Spectrum (b) Layman Crisp Spectrum (c) Crisp QDA

Figure 11: Crisp Spectrums and QDA for Experts and Laymans

With similar patterns across the groups seen with ”warm”, ”bright” and ”clean” and dissimilar patterns visible with ”low”, ”airy”, and ”crisp”, how well the entries agree with each other for a given descriptor can help determine which descriptors vary least from person to person, and thus which have a more uniform perceptual definition. In Table 10. below agreement scores for all common descriptors, calculated using Frechets distance, can be seen. Ordered increasingly, with the most perceptually agreed upon being ”warm”, and least agreed upon being ”pure”.

Descriptor		Agreement	Descriptor		Agreement
1	warm	172.97	21	boxy	1057.51
2	bright	232.12	22	full	1155.24
3	soft	294.01	23	tight	1344.57
4	loud	303.43	24	fat	1433.94
5	crisp	465.62	25	quiet	1435.63
6	clean	546.62	26	light	1639.16
7	harsh	595.84	27	flat	1859.42
8	airy	651.41	28	dull	1934.74
9	muddy	701.18	29	bite	1984.23
10	smooth	761.67	30	good	2241.89
11	sharp	772.81	31	fresh	2391.63
12	low	784.89	32	strong	2547.76
13	thick	855.35	33	tender	2685.74
14	sweet	882.70	34	lively	3001.66
15	bass	918.54	35	vibrant	3484.13
16	nice	931.26	36	twangy	3515.57
17	deep	950.78	37	damped	4779.67
18	dark	984.62	38	big	5724.85
19	aggressive	989.88	39	pure	6402.79
20	punchy	1021.94			

Table 10: Agreement Score of Each Equalisation Descriptor by Frechet Distances

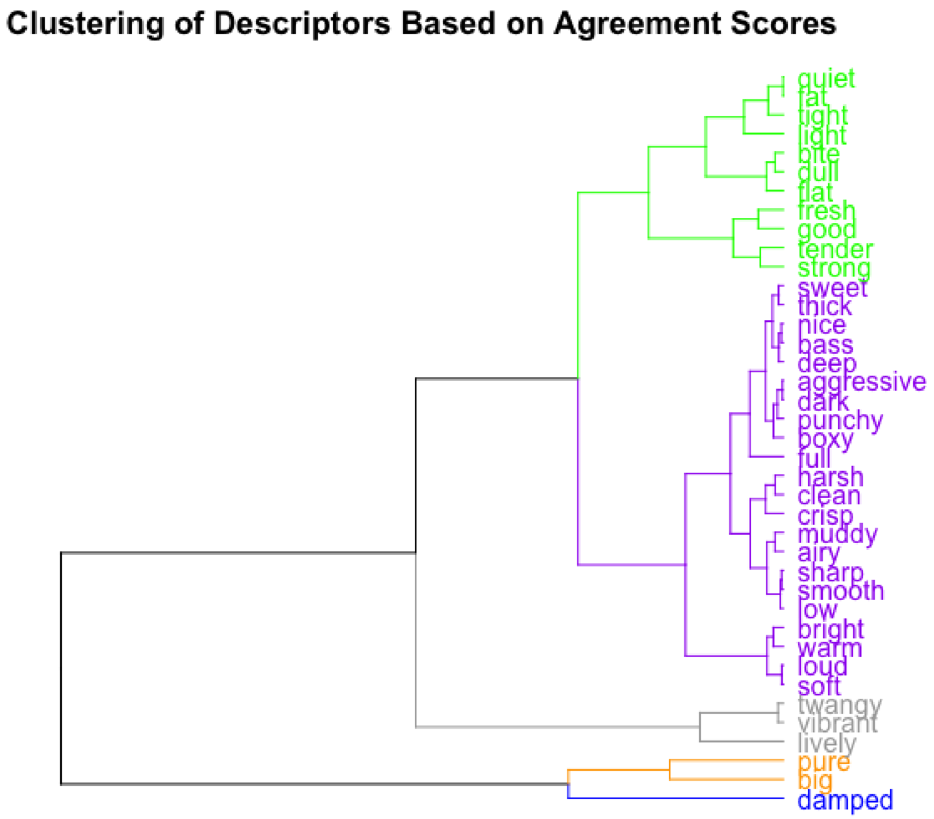
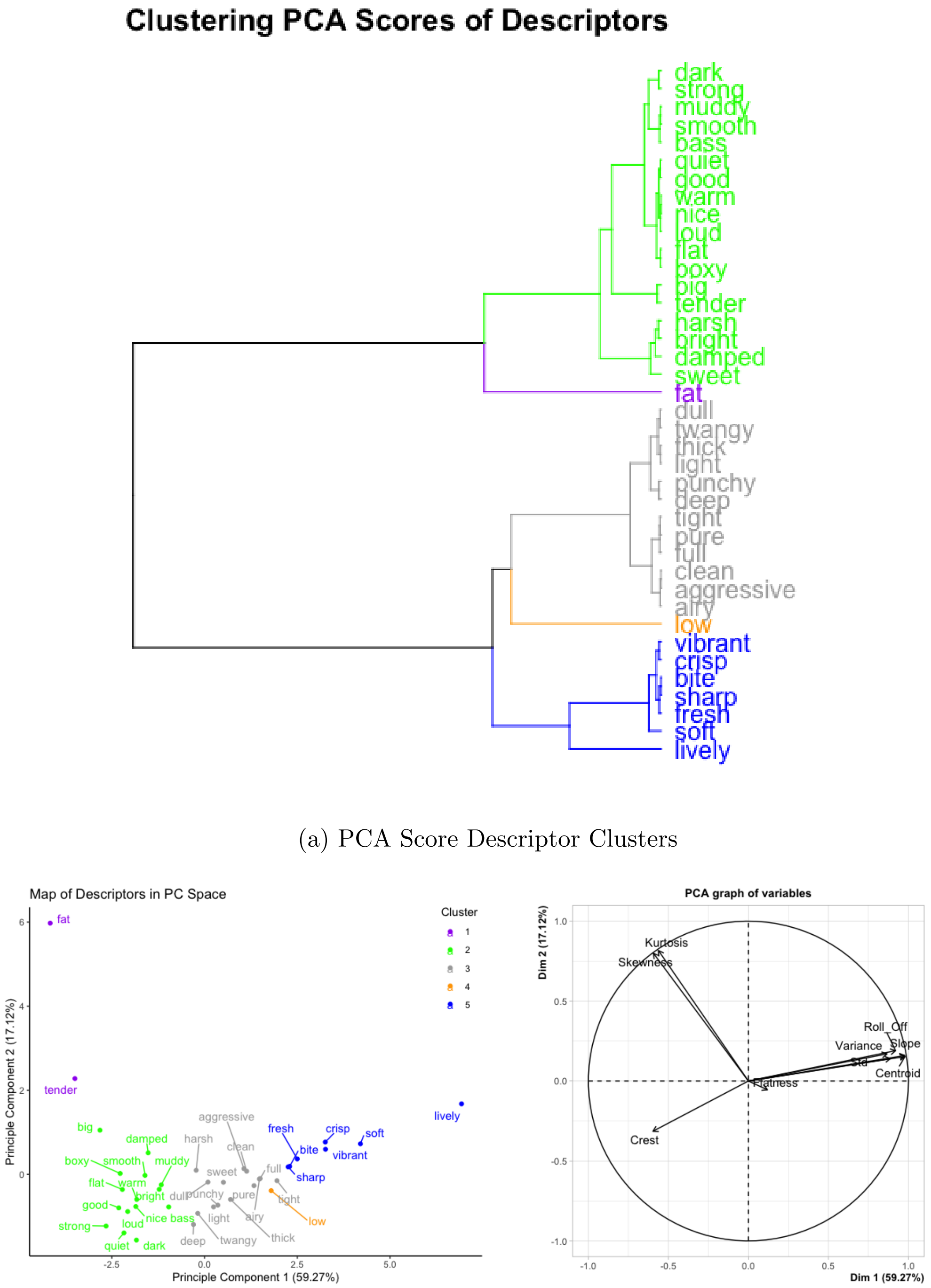


Figure 12: Clustering of Equalisation Descriptors Based on Agreement Score

Clustering by the agreement scores shown in Table 10. produces Figure 12. which is clusters of how non-varying each descriptor is. The clusters are a mixture of synonymy and pairwise dissimilarity, with pairs like "warm" and "bright", "vibrant", "lively", and "twangy", "soft" and "loud", "sharp" and "smooth", and "muddy" and "airy".

Descriptor Feature Space



(a) PCA Score Descriptor Clusters

(b) PCA Clustered Descriptor Distribution

(c) PCA Feature Distribution

Figure 13: PCA of Aggregate Means for Descriptors and Spectral Features

Extracting the spectral features via the spectral equations shown in Table 3. produces nine spectral values for every entry in the dataset. The data is then averaged by descriptor to produce the aggregated values shown in Table 11., that are then used in a Principle Component Analysis. The results of the PCA are shown in Figure 13. with Figure (b) showing the placement of each descriptor in their new shifted space on PC1 and PC2 representing 76.39% of the variation of the original data in just two dimensions. Figure 13. (a) is the respective clustering based of their placement in this feature space, the relationship between Figure 13. (a) and (b) can be seen below in Figure 14.

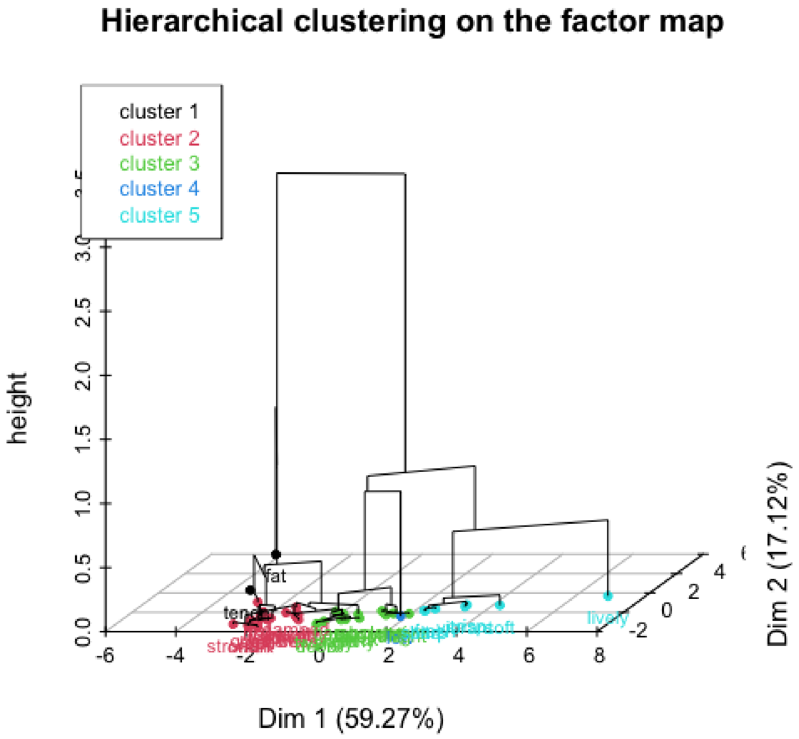


Figure 14: PCA Descriptor Clusters and their Distributions

	Descriptor	Centroid	Variance	Std	Skewness	Kurtosis	Roll Off	Flatness	Crest	Slope
1	warm	2007.41	12282593.36	3357.93	2.70	13.38	755.09	508728673626.31	162.10	-0.00
2	bright	2408.82	12011936.20	3282.44	2.46	12.08	1270.80	1071236041663.46	128.66	-0.00
3	bass	2616.83	17228788.97	3975.38	1.64	7.82	1199.29	1371759.03	199.02	-0.00
4	low	4282.77	25202824.00	4742.70	0.67	1.25	2307.37	21543832234489.04	141.05	-0.00
5	airy	4360.83	23256777.50	4631.85	0.67	0.13	2450.63	688867.40	136.07	-0.00
6	clean	3998.76	24671760.43	4670.56	1.54	4.58	2017.27	1381745.35	127.11	-0.00
7	crisp	5189.20	27517144.38	5014.90	0.81	2.00	3430.49	680223.80	74.08	-0.00
8	full	4035.47	29363564.93	4948.84	0.57	-0.14	1831.84	449900.26	126.93	-0.00
9	boxy	1785.03	11905757.21	3298.09	4.69	28.07	591.72	6434.24	153.87	-0.00
10	thick	3721.06	24854871.35	4758.19	0.49	-0.63	1727.00	290638.50	183.50	-0.00
11	deep	3003.18	26462779.40	4949.18	0.32	-1.28	622.29	248779.91	241.29	-0.00
12	muddy	2630.47	15789118.10	3840.51	3.53	14.84	1138.81	144657.17	178.71	-0.00
13	punchy	3370.50	28420930.52	5115.03	0.57	0.02	804.19	193242.89	195.02	-0.00
14	bite	4543.76	25841110.66	4749.53	0.29	-1.00	2922.19	545671.45	101.62	-0.00
15	aggressive	3981.51	22713196.00	4491.44	1.60	5.66	2001.37	238467.73	114.01	-0.00

Table 11: First 15 Aggregate Means of Spectral Features for Each Common Descriptor

This clustering produced groupings based on similar spectral features and therefore the clusters represent similar perceptual definitions, thus these descriptor clusters are synonymous in nature.

Figure 13. (c) shows how each of the spectral features contribute to each dimension and descriptor, with features such as the centroid, variance, standard deviation, roll off, and slope contributing more to PC1 than to PC2, and features kurtosis and skewness contributing to PC2 more than PC1. The relationship between Figure 13. (b) and (c) can be seen in Table 12. where the descriptors of cluster one, ”fat” and ”tender”, are most effected by skewness and kurtosis, and since the feature cluster means are much greater than their overall means it shows that those descriptors are characterised by high association to those features when compared to other clusters.

Cluster	Descriptors	Feature	Cluster Mean	Overall Mean
1	Fat	Skewness	17.30	2.48
	Tender	Kurtosis	124.56	13.27

Table 12: Feature Association to Clusters

While descriptor frequency occurrence is useful when determining which descriptors are used more prominently by experts or layman, and frequency spectra is better at showing the effect of a given timbre change, it’s the extracted feature space that produces the best dynamic description of a given descriptor or cluster.

4.2 Reverberation

Descriptor Occurrence

Expert			Layman		
Descriptor	n	Decimal	Descriptor	n	Decimal
airy	15	0.07	echo	49	0.11
big	13	0.06	warm	35	0.08
echo	12	0.05	muddy	19	0.04
dreamy	10	0.04	church	18	0.04
hall	8	0.03	big	17	0.04

Table 13: Top 5 Occurring Timbre Descriptors Used by Expert and Layman for Describing Reverberation

		Expert		Layman				Expert		Layman	
Descriptor		n	Decimal	n	Decimal	Descriptor		n	Decimal	n	Decimal
1	airy	15	0.065	4	0.009	6	spacious	7	0.030	11	0.025
2	big	13	0.057	17	0.038	7	soft	6	0.026	6	0.013
3	echo	12	0.052	49	0.109	8	warm	5	0.022	35	0.078
4	dreamy	10	0.043	1	0.002	9	dark	4	0.017	3	0.007
5	hall	8	0.035	6	0.013	10	full	4	0.017	6	0.013

Table 14: First 5 Expert and Layman Common Descriptors for Reverberation

Similar to the equalisation data, after cleaning 230 and 449 entries remained for the expert and layman groups respectively, with the top five occurring descriptors being shown in Table 13. The top 15 descriptors account for 52%, in both groups, of the number of descriptor used meaning that the experts and layman share a similar vocabulary size. To actually examine their shared vocabulary Table 14. and Figure 15. show the top 10 common descriptors and the usage distribution of the 28 shared descriptors, respectively.

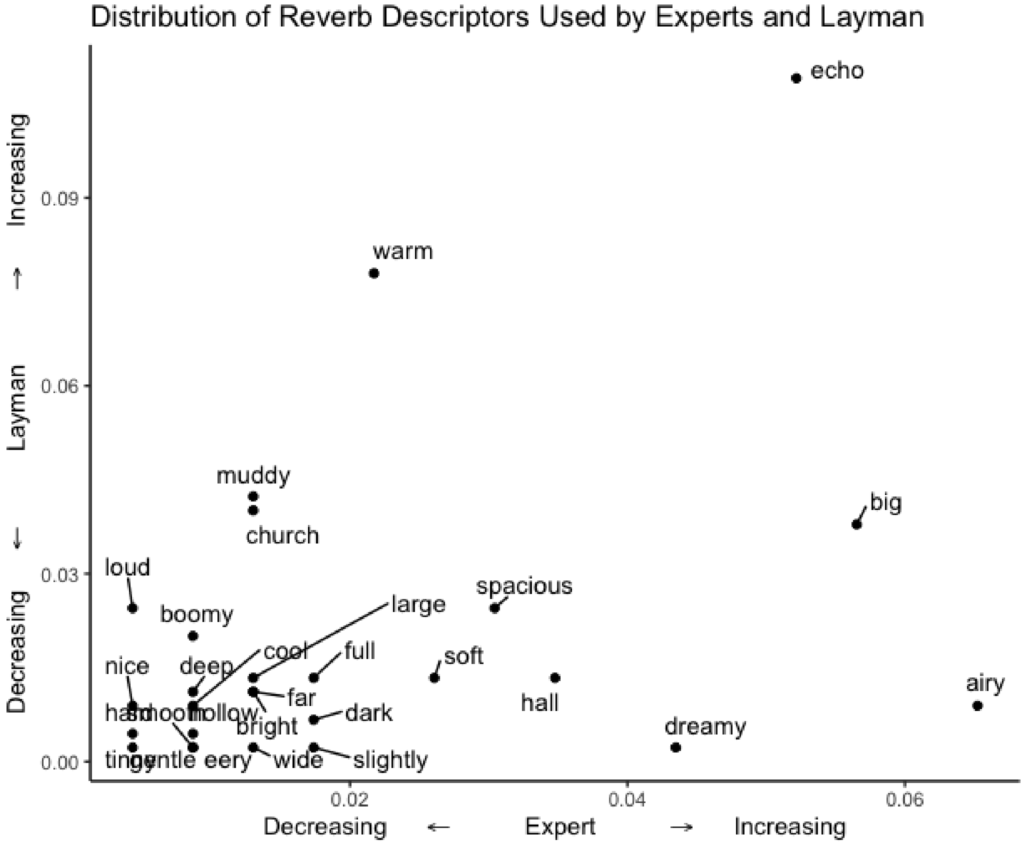


Figure 15: Distribution of Descriptors Used by Experts and Layman for Reverberation

As seen in Figure 15. ”echo” is the most used common descriptor in equal amounts, while descriptors like ”airy” and ”big” are used more frequently by the experts along with the descriptor ”dreamy”. On the other side, descriptors like ”warm”, ”muddy”, and ”church” are more common for the layman to use. The clusters formed by this usage pattern can be seen in Figure 16.

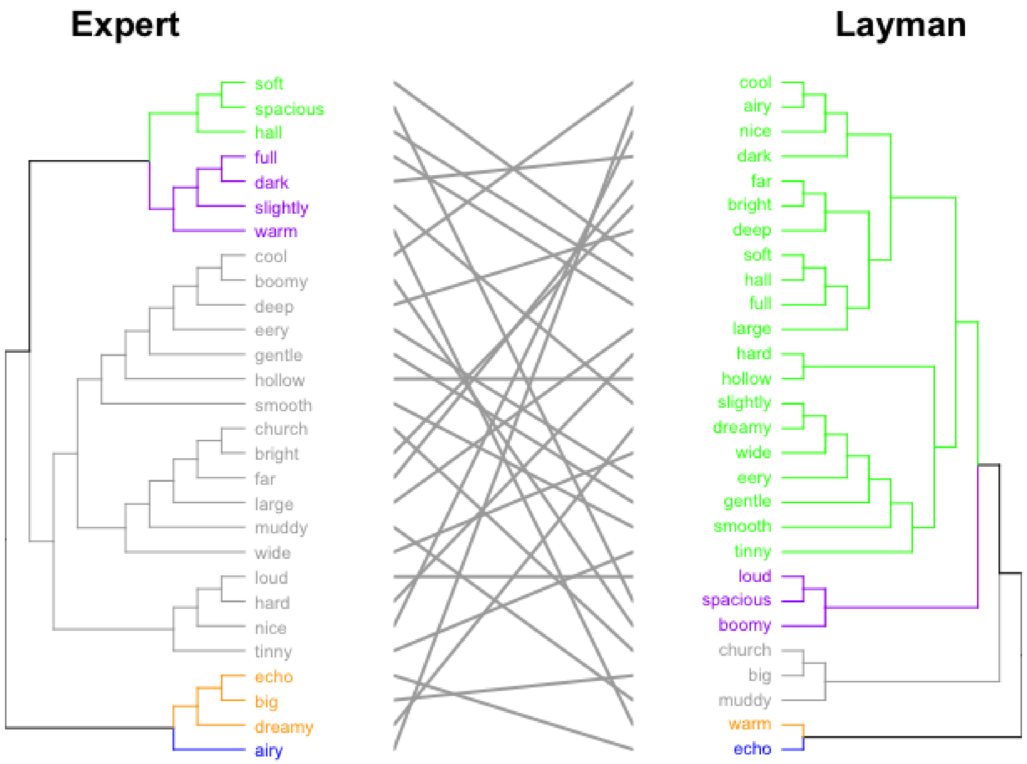
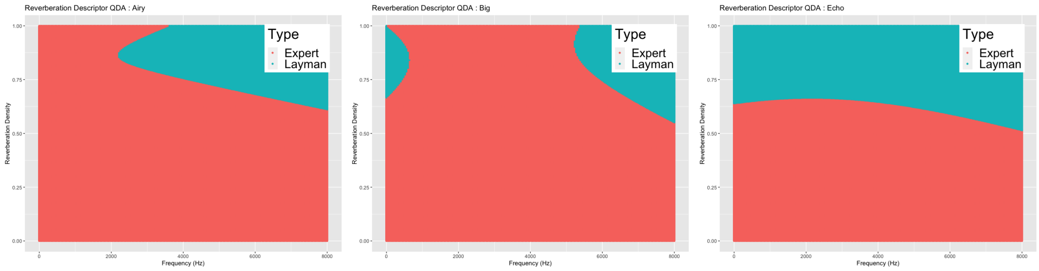


Figure 16: Clustering of Reverberation Descriptors Based on Expert and Layman Usage

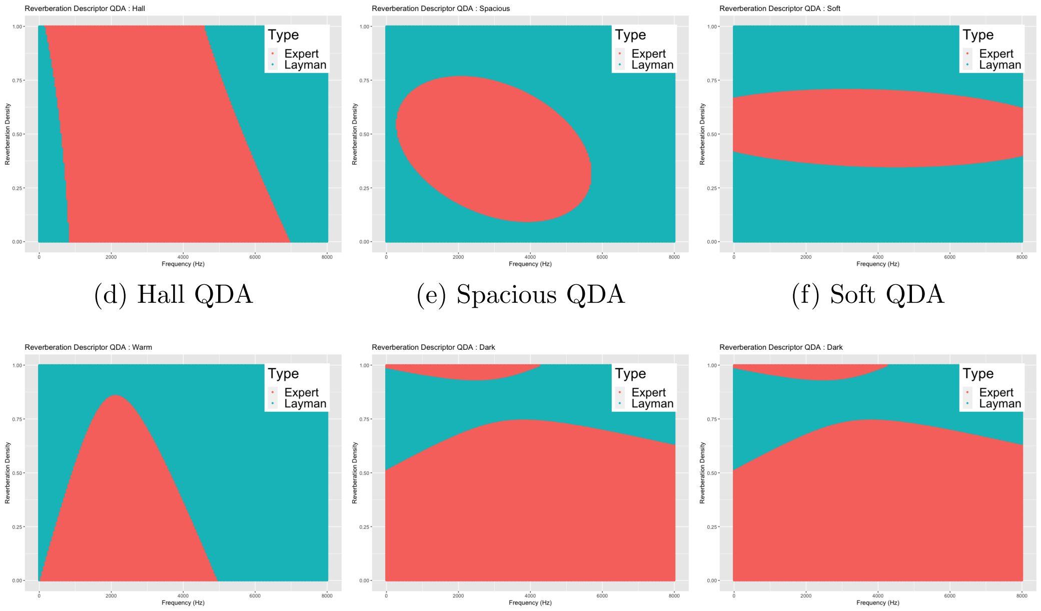
Descriptor Parameter Space



(a) Airy QDA

(b) Big QDA

(c) Echo QDA



(d) Hall QDA (e) Spacious QDA (f) Soft QDA

(g) Warm QDA (h) Dark QDA (i) Full QDA

Figure 17: QDA of Expert and Layman Reverb Descriptors

The QDA's above, again show the frequency space occupation of each group for the first few shared descriptors. Some have very specific frequency definitions like 'spacious" and "soft", which show expert domination in a very contained space with most of the points in this space being close to or part of an expert spectra. Some have equal hold in their frequency space like "echo", "hall", and "warm", with the dominant descriptor definition changing in relation to gain structure for "echo", frequency for "hall", and both for "warm". And some have completely dominated their frequency space like "airy", "big", and "full" with the expert definition being more pervasive. Combining the expert and layman data to determine agreement scores via normalised variance produces the clusters below in Figure 18. The most agreed upon being the descriptor "big", followed by "airy, "muddy", "church", "spacious", and "loud", and the least agreed upon being the furthest away along the tree, "bright", "nice", "large", and "eery".

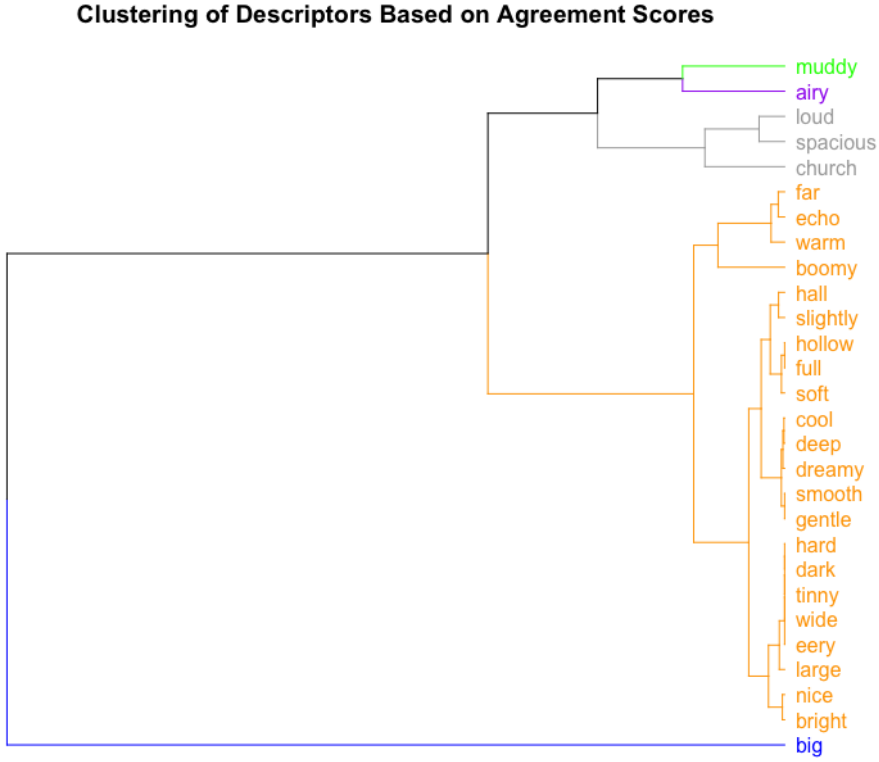
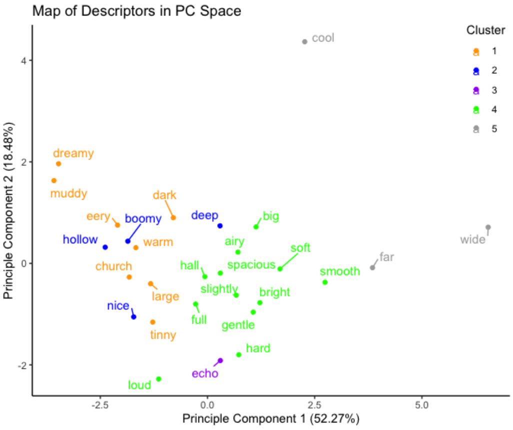
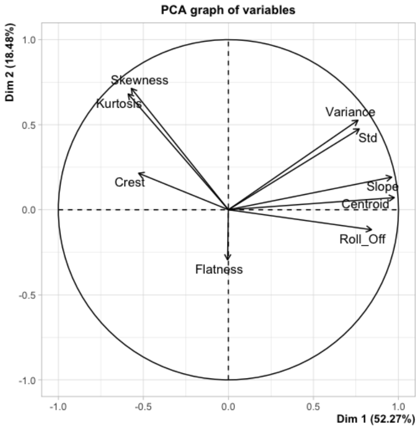


Figure 18: Clustering of Descriptors Based on Agreement Score

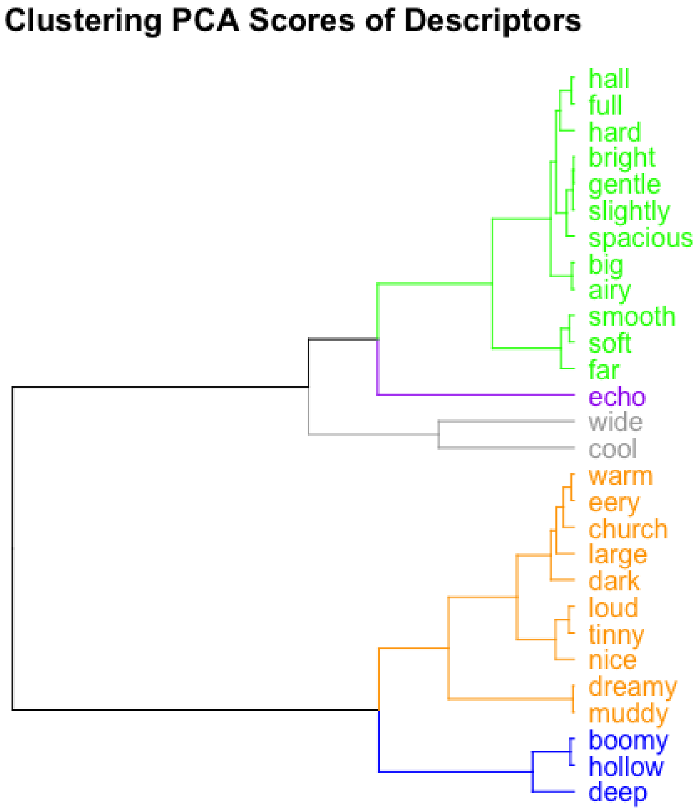
Descriptor Feature Space



(a) PCA Clustered Descriptor Distribution



(b) PCA Feature Distribution



(c) PCA Score Descriptor Clusters

Figure 19: PCA of Aggregate Means for Descriptors and Spectral Features

Again, the spectral features are extracted and reduced to a two dimensional system using PCA that maintains an effective amount of the variation, in this case PC1 and PC2 account for 71%. The results of the PCA are shown in Figure 19., these show that the features centroid, variance, standard deviation, slope, roll off, and crest mostly effect PC1, and skewness, kurtosis, and flatness mostly effect PC2. They also show the distribution of the descriptors in this new space and the clusters formed by feature similarity.

4.3 Effects Comparison

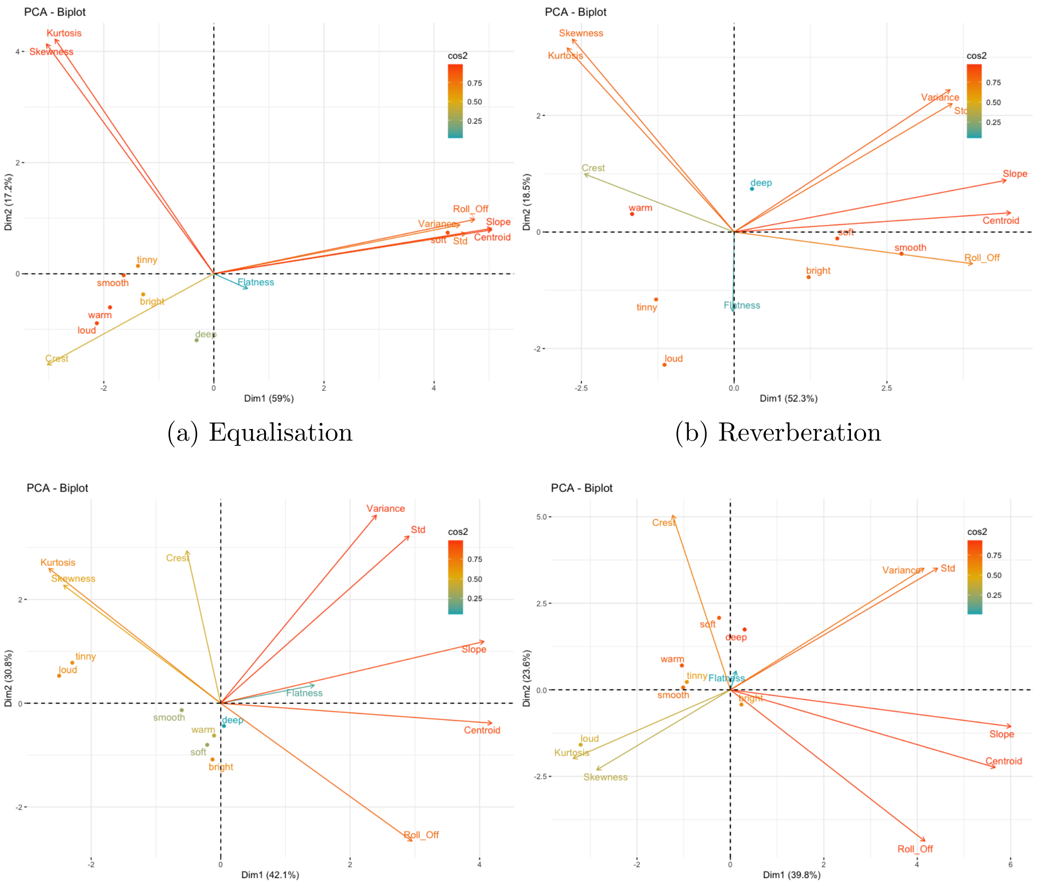
	Descriptor	Equalisation	Reverberation	Compression	Distortion
1	warm	610	40	35	26
2	bright	557	8	10	5
3	tinny	21	2	10	3
4	deep	13	7	5	1
5	smooth	16	3	13	3
6	soft	31	12	16	1
7	loud	27	12	32	2

Table 15: Common Descriptors Across All Timbre Effects

Instead of repeating the same analysis done for equalisation and reverberation for the remaining groups and effects, it's more prudent to simply examine their similarities or lack thereof. As such, the descriptors shared across all effects, equalisation, reverberation, compression, and distortion, and across both groups, experts and layman, are shown in Table 15.

Some potential reasoning as to why these descriptors cross all four effects boils down to how the changes in timbre, caused by each effect, affect our perception of what the descriptor is doing. For instance, descriptors like "warm", "bright", and "tinny" are all strongly associated with equalisation, however when certain changes are made with reverberation, compression, and distortion tools the high frequency components get reduced which can also be percieved as making the sound warmer, brighter, and tinner.

Other descriptors like "smooth", "soft", and "loud" could be associated to all effects since equalisation can produce their spectra by pumping or damping prominent frequencies, reverberation can increase or decrease the density of direct sound, compression can flatten or space out a sound, and distortion can clean up or add noise and dirty a sound. All of which can produce different levels of smoothness, softness, or loudness.



(a) Equalisation

(b) Reverberation

(c) Compression

(d) Distortion

Figure 20: Common Descriptors in Feature Space Across Effects

PCA determines the new shifted dimensions by rotating through all possible axes positions to find the axis that has the most contributions from all descriptors and features, thus maintaining the most variation in a reduced number of dimensions. Figure 20. shows the PCA results for all four effects along with the locations of the shared descriptors in each effects shifted space. The corresponding variations maintained for the two dimensional solution being 76% for equalisation, 71% for reverberation, 73% for compression, and 63% for distortion.

	Centroid	Variance	Std	Skewness	Kurtosis	Roll Off	Flatness	Crest	Slope
Equalisation	18.03	14.12	14.80	6.55	5.88	15.87	0.26	6.47	18.02
Reverberation	20.23	12.33	12.58	6.93	7.39	15.05	0.00	5.93	19.58
Compression	25.42	7.04	10.83	7.40	9.35	13.01	2.83	1.11	23.00
Distortion	22.39	11.99	13.72	5.71	7.90	12.08	0.01	1.07	25.13

(a) Principle Component Dimension 1

	Centroid	Variance	Std	Skewness	Kurtosis	Roll Off	Flatness	Crest	Slope
Equalisation	1.51	1.87	1.27	41.49	43.24	2.32	0.17	6.52	1.61
Reverberation	0.30	16.64	13.58	30.60	27.87	0.83	5.21	2.77	2.20
Compression	0.10	25.23	20.16	10.30	13.26	11.60	0.27	15.64	3.45
Distortion	5.92	14.51	14.53	6.30	4.64	22.50	0.35	29.93	1.32

(b) Principle Component Dimension 2

Table 16: Percent of Contribution of Each Feature per Effect

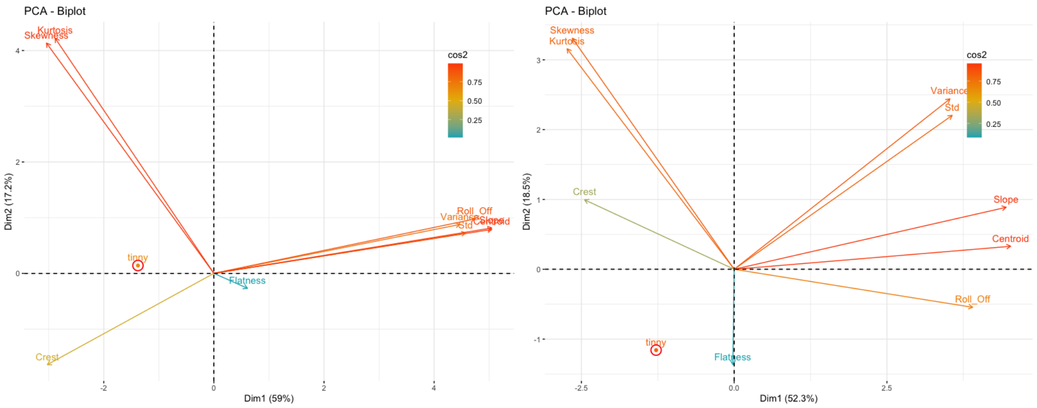
The PC contribution percentages for the spectral features are shown in Table 16. above. It can be seen, that for equalisation, the features centroid, variance, standard deviation, roll off, and slope contribute the most to PC1, and skewness and kurtosis more associated with PC2. A similar breakdown is found for the contributions of reverberation. For compression, PC1's major contribution comes from the features centroid and slope, with PC2's coming from variance and standard deviation. As for distortion, PC1 is similar to the equalisation and reverberation result except for roll off, which is now a larger contributor to PC2 along with the feature crest.

In terms of the descriptors contributions Table 17. shows the breakdown of their PC contributions. From this table it can seen how shared descriptors are affected by the different effects. For example, the descriptor "tinny" contributes two and half times more to compressions PC1 than it's next highest PC, reverberation PC2. Similarly, "soft" contributes over three times as much to equalisations PC1 when compared to it's next highest contribution, reverberation PC1.

	Equalisation		Reverberation		Compression		Distortion	
Descriptor	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
warm	1.72	0.61	2.12	0.20	0.01	0.70	0.39	0.30
bright	0.80	0.23	1.13	1.29	0.02	2.13	0.02	0.11
tinny	0.92	0.03	1.24	2.88	6.93	1.10	0.31	0.03
deep	0.05	2.37	0.06	1.17	0.00	0.35	0.03	1.83
smooth	1.30	0.00	5.70	0.30	0.48	0.03	0.37	0.00
soft	8.70	0.90	2.16	0.03	0.06	1.16	0.02	2.61
loud	2.18	1.32	0.99	11.14	8.21	0.50	3.68	1.52

Table 17: Percent of Contribution of Each Descriptor per Effect

Since the PC's are determined by rotating the axes to produce the maximum variation representation, the contributions of the spectral features and shared descriptors are proportional to their coordinates in PC space. As such, the closer to the origins in Figure 20. the descriptors or features are, the lower their contributions, and thus the lower their representations on those PC's are.



(a) Equalisation

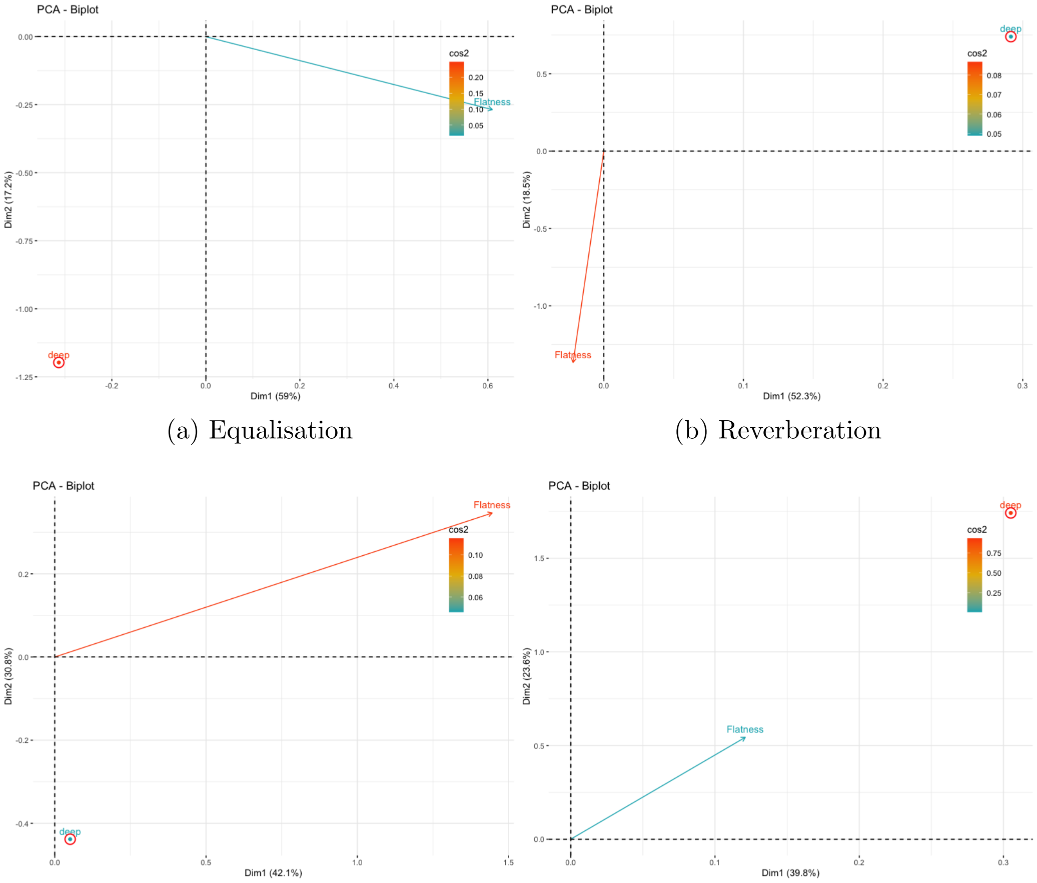
(b) Reverberation

Figure 21: Descriptor Tinny in Feature Space Across Effects

The shared descriptors relate to the spectral features by high or low correlations. For example, the shared descriptors in the equalisation PC space will be correlated based on their locations relative to the features, with descriptors "warm", "bright", "tinny", "smooth", and "loud" being associated with low values of centroid, variance, standard deviation, roll off, and slope, and the descriptor "soft" being associated with high values of those features. This process can be applied to any descriptor in PC space, each with different feature correlation loadings depending on which effect the descriptor was placed in.

For example, Figure 21. shows descriptor "tinny" and the spectral features for equalisation and reverberation. In (a) it can be seen that "tinny" is furthest from the features centroid, variance, standard deviation, roll off, and slope, meaning it will have lower contributions from them in it's frequency spectra, and higher contributions from crest, skewness, and kurtosis. In the reverberation PC space, "tinny" is furthest from variance and standard deviation, meaning it'll be those features that have the lowest contributions in it's respective frequency spectrum.

The features and descriptors in Figure 20. are coloured in accordance with their cos2 values, which is their quality of representation, calculated via squared coordinates. A high cos2 value, a redder colour, indicates a better representation on the graphed PC's, whereas a lower cos2 value, a bluer colour, represents a worse representation on the PC's. For the most part, the extracted features and shared descriptors are represented well by PC1 and PC2 with only flatness being the most under-represented feature across all effects followed up by crest for equalisation and reverberation, crest and skewness for compression, and skewness and kurtosis for distortion, with "deep" being the most under-represented shared descriptor.



(a) Equalisation

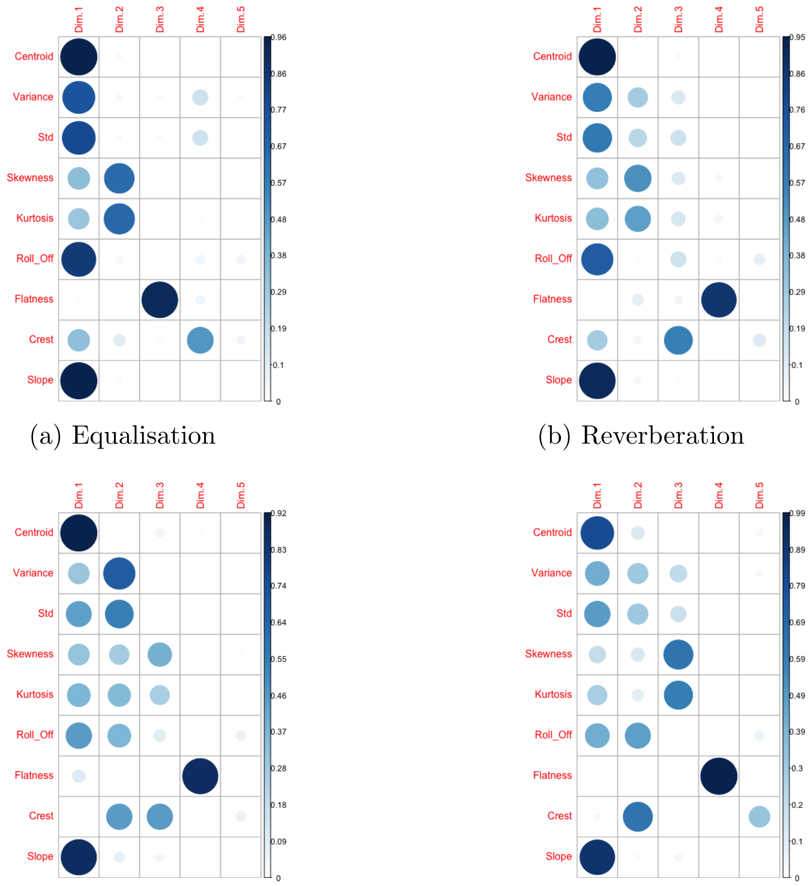
(b) Reverberation

(c) Compression

(d) Distortion

Figure 22: Representation of Descriptors and Features Across Effects

Figure 22. shows a zoomed in look at just descriptor ”deep” and the spectral feature flatness. It should be noted that the cos2 legend went from 0 to 1 in Figure 20. and now goes from 0 to 0.1 in Figure 22. This zoomed in look at the contributions of flatness and ”deep” is to show how close to the origins they are as their contributions are a tenth of other descriptor and feature contributions. Thus, they are not well represented by PC1 or PC2.



(a) Equalisation

(b) Reverberation

(c) Compression

(d) Distortion

Figure 23: Quality of Representation of Variables per Dimension of Each Effect

The breakdown of the quality of representation can be seen in Figure 23., which shows the relative size of cos2 values for each spectral features in every dimension. It can be seen that while PC1 and PC2, the first two dimensions, represent the most features in the most effective way and maintain most of the variation, the latter dimensions do represent some features better. For equalisation, flatness is represented better by PC3, and crest is ever so slightly represented better in PC4. Whereas, in reverberation the representations are flipped, with crest being better in PC3, and flatness in PC4.

	Centroid	Variance	Std	Skewness	Kurtosis	Roll Off	Flatness	Crest	Slope
Equalisation	0.98	0.87	0.89	-0.59	-0.56	0.92	0.12	-0.59	0.98
Reverberation	0.98	0.76	0.77	-0.57	-0.59	0.84	-0.00	-0.53	0.96
Compression	0.97	0.51	0.63	-0.52	-0.59	0.69	0.32	-0.20	0.92
Distortion	0.90	0.65	0.70	-0.45	-0.53	0.66	0.02	-0.20	0.95

(a) Principle Component Dimension 1

	Centroid	Variance	Std	Skewness	Kurtosis	Roll Off	Flatness	Crest	Slope
Equalisation	0.15	0.17	0.14	0.80	0.82	0.19	-0.05	-0.32	0.16
Reverberation	0.07	0.53	0.48	0.71	0.68	-0.12	-0.29	0.21	0.19
Compression	-0.05	0.85	0.76	0.55	0.62	-0.58	0.09	0.67	0.32
Distortion	-0.35	0.55	0.56	-0.37	-0.31	-0.69	0.09	0.80	-0.17

(b) Principle Component Dimension 2

Table 18: Loading Correlations of Each Effect

For compression the cos2 values for crest are similar in both PC2 and PC3, slightly better in PC3 for skewness, and best in PC4 for flatness. Finally, kurtosis and skewness are better when using PC3 and flatness is better when using PC4 in relation to distortion effects. PC5 and above maintain very little significance for this data. In a similar process, it's shown that the descriptor "deep" is best represented by PC3. However, while PC3 and PC4 do contain some information about the effects and the descriptors, PC1 and PC2 are adequate enough to differentiate between the effects, and maps the spectral features well enough to differentiate between descriptors.

Table 18. shows the PC loading correlations of each feature for each effect. From this it can be seen in all cases, as stated above, the feature flatness does not correlate well with PC1 or PC2, nor does it contribute much to those PC's. It can also be seen that features like the variance and the standard deviation have well defined and separated contributions and loadings that can be used to determine which effect is being used.

For PC1 the correlations for variance are 0.8 for equalisation, 0.7 for reverberation, 0.6 for distortion, and 0.5 for compression. So determining the correlation of the variance feature in PC based feature space for a piece of music can help identify which effect might be being applied.

Once a given effect is determined, Figure 20. combined with Figure 19. and Figure 13., along with the associated figures for compression and distortion, the features which best represent a given descriptor can be determined. This allows for a further determination of an appropriate descriptor for a given sound based on it's spectral characteristic features, or the reverse, an appropriate sound for a given descriptor.

Conclusion

It should now be evident why timbre is so difficult to empirically quantify and define in a reasonably scientific manner that produces a universal vocabulary that isn't merely perceptually qualitative descriptors. However, since the auditory experience is a perceptually intuitive response to some sound stimulus, it is inevitable that one persons description of warmth, isn't anothers. And while audio experts have corralled themselves into a succinct list of common descriptors with a much more universally accepted meaning, the audio layman still has a much wider range of descriptors. As seen with equalisation, where the top 15 common descriptors amongst experts accounted for nearly 90% of the expert usage, whereas the layman top 15 only accounted for 35%.

In terms of what the experts and laymen are trying to describe using these particular descriptors, the frequency spectra and QDA's for a selection of shared descriptors across the expert and layman groups are shown in

Figure 6. through to Figure 11. for equalisation and Figure 17. for reverberation. These figures show similarities and dissimilarities, along with, where in frequency space expert or layman descriptor definitions dominate. Furthermore, these figures and the agreement score clusterings shown in Figure 12. for equalisation and Figure 18. for reverberation, show that there are some intuitively quantitative descriptors that share meaning across groups that typifies an effect. For example, "warm" and "bright" for equalisation and "airy" and "big" for reverberation.

Unfortunately, standard frequency space is too restrictive to fully represent the multidimensional nature of timbre. As such, characteristic spectral features were extracted from the descriptors frequency spectra and dimensionally reduced using PCA. This produced a two dimensional solution with maintained variations of 76% for equalisation, 71% for reverberation, 73% for compression, and 63% for distortion. The PC contributions of the spectral features and shared descriptors across all four effects can be seen in Table 16. and Table 17., respectively.

These tables show that the features centroid, variance, standard deviation, roll off, and slope are major contributors to PC1 for equalisation and reverberation, and skewness and kurtosis being bigger factors in PC2 for the same effects. It furthers shows that centroid and slope are more important to compressions PC1, with PC2 being more weighted by the variance and standard deviation. Distortion has a similar PC1 to equalisation and reverberation, with only roll off switching to PC2 along with a major contribution from the spectral crest. The tables also show that the spectral feature flatness, and common descriptor "deep" being under represented by the two dimensional solution, with more of their variation information being stored in latter PC's as seen in Figure 23.

Finally, the loading correlations of the PC's, Table 18., have well separated values for spectral variance and standard deviations such that, identifying which effect is being applied is easier in PC space. For example, the correlations of the variance are 0.8, 0.7, 0.6, and 0.5 for equalisation, reverberation, distortion, and compression, respectively. Once the applied effect is determined, the spectral feature correlations of a specific descriptor within that effects PC space can then also be determined, thus, better defining a given descriptor in a comparative way.

Future Work

In terms of what research comes next, there are a few direct follow-ups to this project and some associated research that can be done in this area. In terms of direct follow-ups, some future analysis can be done into descriptors that are not shared amongst experts and layman, nor do they cross effects boundaries and why some descriptors are shared and others are not.

Another possibility includes an intuitive API for semantic audio processing via a semantic interface. It would work by taking a piece of music and a descriptor, then use its feature space to change the music based on the descriptor by increasing or decreasing the features most associated with that descriptor. It could also use the loadings of features to choose an effect to apply or use the parameter space to allow the user to be either an audio expert or layman.

Analysis into which descriptors are common by genre or musical style is another possible project. Descriptors which invoke some kind of emotive or neurobiological response and why is a related potential research area. And finally, a look at descriptors with some location metadata would be useful for examining if there is a cultural dependency factor involved in how people process music.

References

[1] American Standards Association. (1960). *American Standard Acoustical Terminology.* Technical Report.

[2] Argüelles, M., Benavides, C. and Fernández, I. (2014). *A New Approach to the Identification of Regional Clusters: Hierarchical Clustering on Principal Components.* Applied Economics, 46(21), pp.2511–2519.

[3] Cartwright, M. and Pardo, B. (2013). *Social-EQ: Crowdsourcing an Equalization Descriptor Map.* 14th International Society for Music Information Retrieval.

[4] Cartwright, M., Pardo, B. and Reiss, J. (2014). *Mix-ploration: Rethinking the Audio Mixer Interface.* In Proceedings of the 19th International Conference on Intelligent User Interfaces, pp365–370.

[5] Disley, A.C., Howard, D.M. and Hunt, A.D. (2006) *Timbral Description of Musical Instruments.* In Proceedings of the 9th International Conference on Music Perception and Cognition, Bologna, 2006.

[6] Ferrer, R. and Eerola, T. (2011). *Semantic Structures of Timbre Emerging from Social and Acoustic Descriptions of Music.* EURASIP Journal on Audio, Speech, and Music Processing, 2011(1).

[7] Fitzgerald, R. and Lindsay, A. (2004). *Tying Semantic Labels to Computational Descriptors of Similar Timbres.* International Conference on Sound and Music Computing 2004.

[8] Grey, J. (1977). *Multidimensional Perceptual Scaling of Musical Timbres.* Journal of the Acoustical Society of America, 61(5), pp.1270–1277.

[9] Helmholtz, H. and Ellis, A. (1954). *On the Sensations of Tone as a Physiological Basis for the Theory of Music.* Dover, New York, 2d English Edition.

[10] Howard, D and Angus, J. (2009). *Acoustics and Psychoacoustics.* Focal Press, 4th Edition.

[11] Huber, D. and Runstein, R. (2010). *Modern Recording Techniques.* Focal Press/Elsevier, Amsterdam; Boston, 7th Edition.

[12] Krekovic, G., Poscic, A. and Petrinović, D. (2016). *An Algorithm for Controlling Arbitrary Sound Synthesizers using Adjectives.* Journal of New Music Research, 45(4), pp.375–390.

[13] McAdams, S., Winsberg, S., Donnadieu, S., Soete, G. and Krimphoff, J. (1995). *Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specificities, and Latent Subject Classes.* Psychological Research, 58(3), pp177–192.

[14] Mecklenburg, S. and Loviscach, J. (2006). *subjEQt: Controlling an Equalizer Through Subjective Terms.* Extended Abstracts on Human Factors in Computing Systems, pp1109-1114.

[15] Moravec, O. and Stepanek, J. (2003). *Verbal Description of Musical Sound Timbre in Czech Language.* In Proceedings of the Stockholm Music Acoustics Conference, pages 643–645, Stockholm, Sweden, pp4-5, 2003.

[16] Porter, M.F. (1980). *An Algorithm for Suffix Stripping.* Program, 14(3), pp130–137.

[17] Pratt, R.L. and Doak, P.E. (1976). *A Subjective Rating Scale for Timbre.* Journal of Sound and Vibration, pp45.

[18] Rossing, T., Moore, R. and Wheeler, P. (2002). *The Science of Sound.* Addison Wesley, 3rd Edition

[19] Sabin, A. and Pardo, B. (2009). *2DEQ: An Intuitive Audio Equalizer.* In Proceedings of the 7th ACM Conference on Creativity and Cognition 2009.

[20] Saitis, C. and Weinzierl, S. (2019). *The Semantics of Timbre.* Timbre: Acoustics, Perception, and Cognition, pp.119–149.

[21] Saitis, C., Fritz, C. and Scavone, G.P. (2017). *Perceptual Evaluation of Violins: A Psycholinguistic Analysis of Preference Verbal Descriptions By Experienced Musicians.* Journal of the Acoustical Society of America, 141, pp2746–2757.

[22] Sarkar, M., Vercoe, B. and Yang, Y. (2007). *Words that Describe Timbre: A Study of Auditory Perception Through Language.* In Proceedings of Language and Music as Cognitive Systems Conference 2007.

[23] Sayago, A., Asuero, A. and Gonzalez, G. (2006). *The Correlation Coefficient: An Overview.* Critical Reviews in Analytical Chemistry. 36(1). pp41-59.

[24] Seetharaman, P. and Pardo, B. (2014). *Crowdsourcing a Reverberation Descriptor Map.* Proceedings of the ACM International Conference on Multimedia 2014.

[25] Seetharaman, P. and Pardo, B. (2014). *Reverbalize.* Proceedings of the ACM International Conference on Multimedia 2014.

[26] Seetharaman, P. and Pardo, B. (2016). *Audealize: Crowdsourced Audio Production Tools.* Journal of the Audio Engineering Society, 64(9), pp.683–695.

[27] Siedenburg, K., Fujinaga, I. and McAdams, S., 2016. *A Comparison of Approaches to Timbre Descriptors in Music Information Retrieval and Music Psychology.* Journal of New Music Research, 45(1), pp.27-41.

[28] Solomon, L.N. (1958). *Semantic Approach to the Perception of Complex Sounds.* Journal of the Acoustical Society of America, 30, pp421–425.

[29]Soraghan, S., Faire, F., Renaud, A. and Supper, B. (2018). *A New Timbre Visualization Technique Based on Semantic Descriptors.* Computer Music Journal, 42(1), pp.23–36.

[30] Stables, R., Enderby, S., De Man, B. and Reiss, J.D. (2014). *SAFE: A System for the Extraction and Retrieval of Semantic Audio Descriptors.* In Proceedings of the 15th International Society for Music Information Retrieval Conference 2014.
[31] Stables, R., De Man, B., Enderby, S., Reiss, J.D., Fazekas, G. and Wilmering, T. (2016). *Semantic Description of Timbral Transformations in Music Production.* In Proceedings of the 2016 ACM on Multimedia Conference 2016.
[32] Stahl, D., Leese, M., Landau, S. and Everitt, B. (2011). *Cluster Analysis.* 5th Edition. Wiley.
[33] Stepanek, J. (2006). *Musical Sound Timbre: Verbal Descriptions and Dimensions.* In Proceedings of the 9th International Conference on Digital Audio Effects, Montreal, Canada, 2006.
[34] Sundaram, S and Narayanan, S. (2007).*Analysis of Audio Clustering Using Word Descriptions.* In Proceedings of the International Conference of Acoustics, Speech and Signal Processing 2007.
[35] Toulson, E. (2003). *A Need for Universal Definitions of Audio Terminologies and Improved Knowledge Transfer to the Audio Consumer.* In Proceedings of The Art of Record Production Conference 2003.
[36] Turk, S., Blas, M. and Pohar, M. (2004). *Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study.* Advances in Methodology and Statistics. 1(1). pp143-161
[37] Usman, A., Shahzad, A., Javed, F., Atif, M. and Abbas, R. (2017). *Using PCA and Factor Analysis for Dimensionality Reduction of Bio-informatics Data.* International Journal of Advanced Computer Science and Applications. 8(5).
[38] Vlaj, D., Kacic, Z. and Kos, M. (2013). *Acoustic Classification and Segmentation Using Modified Spectral Roll-off and Variance-based Features.* Digital Signal Processing. 23(2), pp659-674.
[39] Von Bismarck, G. (1974) *Timbre of Steady Tones: A factorial Investigation of its Verbal Attributes.* Acustica, 30, pp146–159.
[40] Wallmark, Z. (2018). *A Corpus Analysis of Timbre Semantics in Orchestration Treatises.* Psychology of Music, 47(4), pp.585–605.
[41] Wallmark, Z. (2019). *Semantic Crosstalk in Timbre Perception.* Music and Science, 2, pp.1-18.
[42] Zacharakis, A., Pastiadis, K., Papadelis, G., Reiss, J. (2011). *An Investigation of Musical Timbre: Uncovering Salient Semantic Descriptors and Perceptual Dimensions.* 12th International Society for Music Information Retrieval Conference, pp.807–812.

[43] Zacharakis, A., Pastiadis, K., Papadelis, G., Reiss, J. (2012). *Analysis of Musical Timbre Semantics through Metric and Non-Metric Data Reduction Techniques.* 12th International Conference of Music Perception and Cognition.

[44] Zheng, T., Seetharaman, P. and Pardo, B. (2016). *SocialFX: Studying a Crowdsourced Folksonomy of Audio Effects Terms.* In Proceedings of the 2016 Association for Computing Machinery on Multimedia Conference 2016.