Proceedings of the 2nd AES Workshop on Intelligent Music Production, London, UK, 13 September 2016

DESCRIPTOR SUB-REPRESENTATIONS IN SEMANTIC EQUALISATION

Spyridon Stasis, Jason Hockman, Ryan Stables

Digital Media Technology Lab
Birmingham City University
{spyridon.stasis,jason.hockman,ryan.stables}@bcu.ac.uk

ABSTRACT

In semantic equalisation, descriptions of audio transformations can be used to control low-level audio effect parameters. In this paper, we explore sub-representations of these descriptions in order to suggest more contextually relevant processing parameters to users, based on external influence. We propose a methodology for finding sub-representations, and an intuitive low-dimensional interface, which can be used to recommend equalisation curves based on proximity to cluster centroids.

1. OVERVIEW

Semantically-informed equalisation involves learning some relationship between a subjective description of musical timbre and a set of audio features, such that a complex parameter space can be controlled using an intuitive low-dimensional interface. A common way to do this is to collect descriptive metadata from audio engineers applying equalisation to a range of audio signals, then to create a map between the two spaces through a process of abstraction. Recently, this has been implemented through the use of dimensionality reduction applied to a parametric equaliser (EQ) [1], or via multiple regression applied to bands of a graphic EQ [2].

In music production, equalisation has a common vocabulary, in which some of the more frequently used descriptors include *warm*, *bright*, *air*, *crisp* and *thin* [3]. Whilst these terms provide meaningful representations of EQ parameter spaces, there is often high within-term variance in the data, suggesting disagreement amongst participants.

In this paper, we attempt to explain this variance by identifying *sub-representations* of semantic descriptors. These are clusters of data points, that can be explained by some external influence. We propose that by identifying the sub-representation a user is trying to achieve, we can provide a more contextually relevant parameter space for the descriptor they are working with. This is done through an intuitive predictive interface that can be navigated in two dimensions.

2. METHODOLOGY

To identify descriptor sub-representations, we apply clustering to a dataset of annotated equalisation settings. Clusters are found within each individual term, for settings described as both *warm* and *bright*. We then measure the saliency of the clusters using a number of metrics and develop an interface for cluster navigation.

2.1. Dataset

The dataset used for the experiment is taken from the SAFE EQ [4], in which 582 entries were labelled as *warm* and 531 entries were labelled as *bright*. The annotated settings were collected from audio effects plugins that operate within a digital audio workstation, from a corpus of anonymous users. For each entry into the dataset, the setting has a string of semantic descriptors, a feature vector containing over 100 audio features per frame extracted before and after processing, and a parameter space vector that describes the gain, bandwidth and centre frequency for each biquad filter (1x lo-shelf, 3x peak, 1x hi-shelf) in the EQ.

2.2. Clustering

To find sub-representations we first apply dimensionality reduction to the parameter space using a stacked autoencoder, allowing us to reduce the 13-dimensional set of EQ controls to a navigable 2-dimensional space. K-means is then applied to the low-dimensional space in order to find clusters of entries. The optimal number of clusters is set to 3 for both *warm* and *bright* by maximising the group separability using a silhouette score.

To ensure the data is capable of forming reliable partitions, we measure the Hopkins Statistic of clustering tendency, which estimates the likelihood of data points being sampled from a non-uniform distribution by comparing them with randomly sampled values from the low-dimensional space. Once k-means has been applied, we then measure the saliency of the sub-representations using two metrics. *Ideal Correlation* (IC): measures the coherence between a similarity matrix of the points in the dataset with a matrix of binary values, where cells are set to 1 if points are from the same cluster, and 0 if not. *Average Silhouette Score* (AS): measures the compactness and isolation of clusters by using cohesion (c) and separation (s), where:

$$AS = (S_i - C_i)/max(S_i, C_i) \quad (1)$$

To evaluate the influence of external data on the clusters, we measure the divergence between audio feature distributions within each cluster, before and after audio processing has been applied. This is done using Kullback-Leibler Diverge (KLD), where the target distribution (P) is the feature set after processing, and the approximation distribution (Q) is the input feature set.

3. RESULTS

The data exhibits strong clustering tendency, with a Hopkins Statistic of 0.561 (SD: 0.043) for *warm* and 0.543 (SD: 0.027) for *bright*. This suggests the formation of sub-representations is plausible given the organisation of settings in reduced-dimensionality space. The validity of the sub-representations after clustering is shown in Table 1, where strong positive results are seen for both descriptors. This suggests the formation of reliable clusters after k-means has been applied.

	Ideal Correlation	Silhouette Score	μ
Warm	0.6858	0.4685	0.5771
Bright	0.6148	0.4302	0.5224

Table 1: Cluster validity metrics for both descriptors after k-means

KLD is applied to the feature distributions, before and after audio processing and the features from each cluster are ranked. Only features that were significantly higher ($p > .05$) than the distribution mean were included. Within the *warm* data, Smoothness (10.01) and Tonality (5.37) were salient for cluster 1, and Spectral Flatness (14.62) was salient for cluster 3. Within the *bright* data, Smoothness (12.6) and MFCC 9 (12.09) were salient for cluster 1, Spectral Flatness (16.8) was salient for cluster 2, and MFCC 12 (11.71) was salient for cluster 3. This suggests the groups can be represented using changes in external feature data.

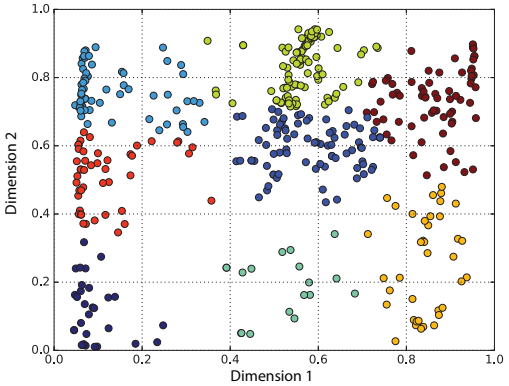
4. INTERFACE

The interface (as shown in Figure 1(a)) allows users to navigate the sub-representations, benefitting from recommended settings in real-time. Modifications to the equalisation parameters can be applied in either low- or high-dimensional space, where the relevant sub-representation is found by minimising the euclidean distance between the user-input and each of the cluster centroids in 2-dimensional space.

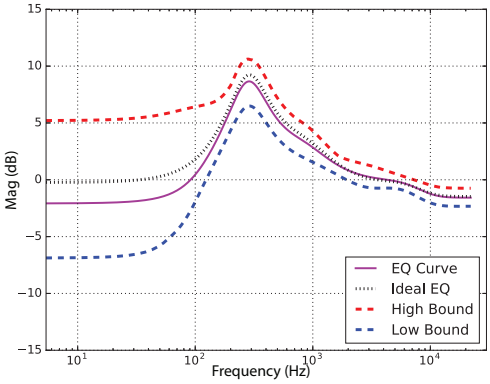
The frequency analyser (Figure 1(b)) provides feedback about (1) the current curve, (2) the boundaries of the current sub-representation and (3) the ideal EQ curve, given the current sub-representation. To derive bounding curves, all of the points in the 2-dimensional cluster are mapped to the parameter space using the decoder layers in the auto-encoder, then the minimal and maximal values are selected from each parameter. To find the ideal EQ curve, the centroid of the current cluster is used as the input to the decoder, resulting in a 13-dimensional vector of EQ parameters.

5. CONCLUSION

We identify sub-representations in a dataset of semantically annotated equalisation data. This is done by applying clustering in reduced dimensionality space and applying clustering tendency measures to measure salience. We evaluate the extent to which additional metadata (audio features) captured using the SAFE plugins[1] can describe the clusters and find that Smoothness, Spectral Flatness and MFCCs score particularly highly with a number of clusters. We conclude by presenting an interface that allows users to explore the sub-representations in a low-dimensional space whilst making recommendations based on cluster centroids.



(a) Descriptor entries mapped into a navigable 2-dimensional space with 8 clusters



(b) The corresponding equalisation interface

Figure 1: The 2-part Interface for cluster navigation

6. REFERENCES

[1] S. Stasis, R. Stables, and J. Hockman, "A model for adaptive reduced-dimensionality equalisation," in *DAFx*, 2015.

[2] M. B. Cartwright and B. Pardo, "Social-EQ: Crowdsourcing an equalization descriptor map," in *ISMIR*, 2013.

[3] R. Stables, S. Enderby, B. De Man, T. Wilmering, G. Fazekas, and J. D. Reiss, "Semantic description of timbral transformations in music production," in *ACM Multimedia*, 2016.

[4] R. Stables, S. Enderby, B. De Man, G. Fazekas, and J. D. Reiss, "SAFE: A system for the extraction and retrieval of semantic audio descriptors," in *ISMIR*, 2014.

[1] Available via `http://www.semanticaudio.co.uk`