SAFE: A system for the extraction and retrieval of semantic audio descriptors

Stables, R; Enderby, S; Man, BD; Fazekas, G; Reiss, JD; 15th Int. Society for Music Information Retrieval Conference (ISMIR-14)

SAFE: A SYSTEM FOR THE EXTRACTION AND RETRIEVAL OF SEMANTIC AUDIO DESCRIPTORS

Ryan Stables, Sean Enderby
DMT Lab
Birmingham City University
ryan.stables@bcu.ac.uk
sean.enderby@mail.bcu.ac.uk

Brecht De Man, György Fazekas, Joshua Reiss
Centre for Digital Music
Queen Mary University of London
b.deman@qmul.ac.uk
g.fazekas@qmul.ac.uk
joshua.reiss@qmul.ac.uk

ABSTRACT

We present an overview of the Semantic Audio Feature Extraction (SAFE) Project, a novel data collection architecture for the extraction and retrieval of semantic descriptions of musical timbre, deployed within the digital audio workstation. By embedding the data capture system into the music production workflow, we are able to maximise the return of semantically annotated music production data, whilst mitigating against issues such as musical and environmental bias. Users of the plug-ins are able to submit semantic descriptions of their own music, whilst utilising the continually growing collaborative dataset of musical descriptors. In order to provide more contextually representative timbral transformations, the dataset is partitioned using metadata, captured within the application.

1. INTRODUCTION

In music production, audio effects are controlled using parameters that interface with statistical, low-level properties of the audio signal. In order for users to achieve the desired timbral transformation, they need to be familiar with this low-level parameter space, often requiring extensive training. Generally, the desired characteristics of sounds can be described using semantic or high-level descriptors such as bright, warm or fuzzy. These perceptual descriptors are not always defined mathematically, and often have some nonlinear relationship to the audio effect's parameter space. It is therefore useful to interpret the relationship between these low-level parameter spaces and high-level descriptions of musical timbre, allowing us to design audio effects that can be controlled with semantically understandable parameters [2]. For this to take place, it is important to compile a dataset of timbral transformation data, complete with reliable semantic annotation.

Musical semantics data is commonly collected using controlled listening tests, with predefined samples and descriptors [4]. The sample sizes are generally small (10-50 subjects), and the descriptive terminology is often limited to a corpus of terms provided by the researcher. Both of these factors suggest that the results may not be scalable to large user-groups, and furthermore dismiss the influence of contextual factors such as genre, musical instrument and location. In this work we present an overview of the Semantic Audio Feature Extraction (SAFE) Project, a system for the extraction and retrieval of semantic audio descriptors from within the music production workflow.

1.1 The SAFE Project

The SAFE project[1] comprises a suite of Digital Audio Workstation (DAW) plug-ins, which encourage the annotation of parameter states during the production process. Users are presented with a free-text field on the UI, allowing them to input multiple text labels simultaneously. As the descriptors are entered into the system, they are uploaded anonymously to the server along with a time-series matrix of audio features, a static parameter space vector and a selection of metadata tags. To motivate the user base to provide this data, the plug-ins allow them to load semantic profiles that are stored on the server, from within the same interface (see Fig. 1).

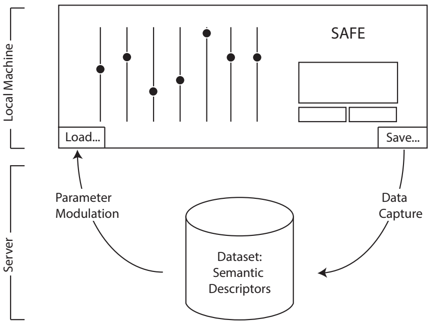
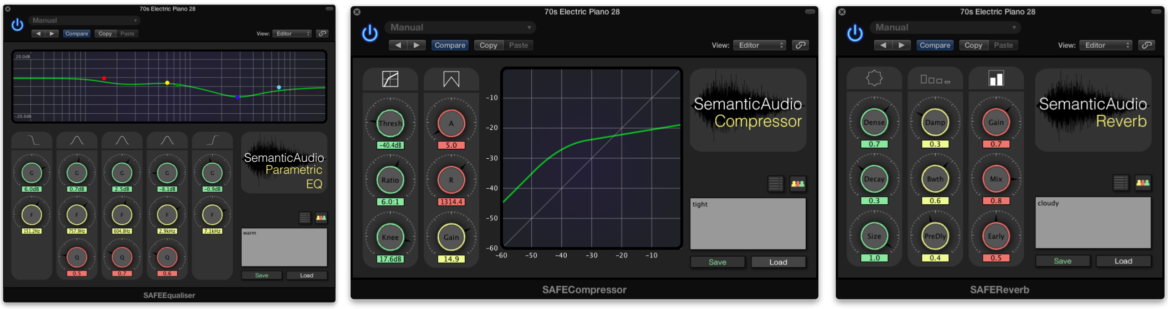


Figure 1. A schematic representation of the plug-in architecture, providing users with load and save functionality.

[1] Plug-ins can be downloaded from http://www.semanticaudio.co.uk



G 6.0dB, F 151.2Hz
G 0.7dB, F 757.9Hz, Q 0.5
G 2.5dB, F 604.8Hz, Q 0.7
G -8.1dB, F 2.9kHz, Q 0.6
G -0.9dB, F 7.1kHz
warm

Thresh -40.4dB, Ratio 6.0:1, Knee 17.6dB
A 5.0, R 1314.4, Gain 14.9
tight

Dense 0.7, Decay 0.3, Size 1.0
Damp 0.3, Bwth 0.6, PreDly 0.4
Gain 0.7, Mix 0.8, Early 0.5
cloudy

(a) Equaliser

(b) Compressor

(c) Algorithmic Reverb

Figure 2. Graphical user interfaces of the equalisation, compression and reverberation plug-ins, with example descriptors.

2. SYSTEM OVERVIEW

2.1 Digital Audio Effects

The system currently consists of four audio effect plug-ins, packaged for VST, Audio Unit, and LV2 formats; an amplitude distortion effect with tone control, an algorithmic reverb based on the figure-of-eight technique proposed by Dattorro [3], a dynamic range compressor with attack and release parameters, and a five band parametric EQ with three peaking filters and two shelving filters. All visible parameters are included in the parameter space vector, and can be modulated via the text input field. The UI for three of the plug-ins is illustrated in Fig. 2, along with examples of current parameter settings.

2.2 Analysis Framework

To represent the signal associated with each descriptor, we store an $N \times M$ matrix of audio features, where N is the number of frames and M is the number of audio features. These are extracted using the LibXtract library [1], an audio feature extraction framework which includes around 40 different features, taken from 10 different input representations. These include temporal features such as log attack time, spectro-temporal features such as spectral centroid and kurtosis, and augmented feature vectors such as MFCCs and ERBs. An analysis block size of N values is used for each feature vector, where N can be read from the host or set empirically. In order to represent the timbral transformation imposed by the audio effect, the feature matrix is computed before and after the processing occurs and differential measurements are taken.

Along with the feature matrix, a $1 \times P$ parameter vector is stored, where P is the number of UI parameters. Based on the current system, P ranges from 6 to 13 parameters. Furthermore, an optional metadata stage is provided to store user and context information. This allows users to disclose their age, location, production experience, the genre of the song and musical instrument of the track. These were deemed to be statistically significant factors in the variance of semantic terminology between different user groups.

2.3 Parameter Modulation

To modulate the UI parameters, users can search for existing semantic profiles on the server and apply them to their own audio signals. Each semantic profile is updated in real-time, meaning they change dynamically based on new input to the server. To provide users with a more reliable representation of their semantic term, the terms are hierarchically partitioned (when possible) into metadata categories. This means that users are able to load instrument, genre and location-specific terms, as opposed to generic terms that cover a wide range of musical conditions. Additionally, transformations from nonlinear effects are applied relative to the signal's RMS to ensure timbral modifications are applied independently of signal level.

2.4 Missing Data Approximation

Due to the nature of the system, users frequently omit optional metadata tags, providing only audio data, the parameter space and text descriptors. In these cases, we are able to approximate missing data using a number of techniques, thus improving the reliability of the semantic parameter settings. Here, the user's location is approximated by storing the geolocation data relating to the IP address and both the musical instrument and genre tags are estimated using an unsupervised machine learning algorithm, applied to a reduced-dimensionality representation of the audio feature set.

3. REFERENCES

[1] Jamie Bullock. Libxtract: A lightweight library for audio feature extraction. In *Proceedings of the International Computer Music Conference*, volume 43, 2007.

[2] Mark Cartwright and Bryan Pardo. Social-eq: Crowdsourcing an equalization descriptor map. In *ISMIR*, pages 395–400, 2013.

[3] Jon Dattorro. Effect design, part 1: Reverberator and other filters. *Journal of the Audio Engineering Society*, 45(9):660–684, 1997.

[4] Alastair Disley and David Howard. Spectral correlates of timbral semantics relating to the pipe organ. *Speech, Music and Hearing*, 46:25–39, 2004.