# An analysis and review of feature extraction and dimensionality reduction techniques for the visual exploration of large audio collections.

Engineering thesis


**Author**: Łukasz Piotrak (Student ID: s18002)
**Supervisor**: -
**Co-Supervisor**: -

Department of Information Technology,
Polish-Japanese Academy of Information Technolog

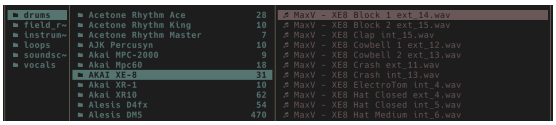
January 29, 2020

# Abstract

The current paradigm for the storage and classification of audio data is that of an attribute ontology; files are grouped together into classes (usually by directory name) i.e. "Drums", "Field recordings" and assigned a range of attributes (by means file name or tags) i.e. "Low", "Kick", "Noisy" etc.

There are a lot of problems with this approach, e.g. The files might mislabeled or the labels might not be an adequate description of the sound. Moreover, relationships between samples are unable to be sufficiently represented (Two audio samples might be in separate branches of the taxonomy but be perceptually similar).

Figure 1: The typical way of exploring files.




However, this method of browsing highlights two crucial attributes of an effective file browser: Firstly, The samples must be visually represented in order to be navigable. Secondly files must be available for immediate inspection when identified.

A more natural representation of the collection might be achieved focusing on perceptual features (such as Timbre), which are inherent to the data itself, and can be calculated using various information retrieval techniques.

These features can be represented in a variety of ways, some of the most popular being Short-term-Fourier-Transform (STFT) and Mel-Frequency-Cepstral-Coefficients (MFCC), however after extracting such features we are still left with the problem of visually representing the sound library.

Recent advances in dimensionality reduction techniques, such as T-sne (van der Maaten & Hinton, 2008) and Umap (McInnes, Healy, & Melville, 2018) have enabled the 2-dimensional visualization of high-dimensional data in an efficient and useful manner, highlighting similarities and relationships between data points. These algorithms are most prominently used to visualize high-dimensional representations of data learnt by deep neural networks, enabling a higher level of intuition when tuning hyper parameters for the algorithms.

However, in this paper I will use these algorithms as a method of reducing the high-dimensional representations of features extracted from audio files into the 2d domain, where they can be intuitively explored.

# Contents

# 1 Algorithm Overview

In order to arrive at a an effective algorithm for visualization I will follow a pipeline similar to the one presented in (Sarwate, 2017)

## 1.1 Dataset

Quite a few labeled audio datasets exist. The most interesting and expansive of these is the google-backed audioSet. This dataset consists of youtube clips labeled in an expansive ontology. However, the dataset is focused more towards classification and is too expansive (over 2.1 mln samples!) for visualization

In the end I have settled on the Medley-solo dataset, (Lostanlen, Cella, Bittner, & Essid, 2019) over 21,000 samples of solo instrument performances divided into 8 classes. Each sample in this dataset has a fixed length. However, my aim is to test the model with samples of differing lengths. Because of this have chosen to supplement the training of the model with the IRMAS dataset (Bosch, Fuhrmann, & Herrera, 2014). It also contains samples of performances classified into 11 classes. However, it differs from Medley in two main aspects:

Firstly, the performances are not solo but in an ensemble. The most prominent instrument is labeled, along other instruments present in the recording. Second, The samples are of different length, which will enable the validation of the method for homogenizing samples of different length.

## 1.2 Feature Extraction & Dimentionality Reduction

A two step process is adopted. First, feature extraction is applied on the data set, followed by one or several dimensionality reduction steps.

A collection of feature extraction methods will be used:

- Short-Term Fourier Transform
- Mel-Frequency Cepstral Coefficients
- Various Music Information Retrieval Features as proposed by Hantrakul et al. (Sarwate, 2017).

To collapse features extracted from samples of different length into a uniform representation the following process will be applied as presented as presented in this blog post by Leon Fedden (Fedden, 2017)

- Mean for each feature dimension
- Standard deviation for each feature dimension

- Mean first order difference for subsequent features frames.

Following (Sarwate, 2017) k-means clustering will be applied to each feature set as a preliminary step to verify the quality of ground truth labels and feature extraction on high-dimensional data.

The following dimensionality reduction techniques will be applied for each feature set:

- Principal Component Analysis
- T-sne
- UMAP

## 1.3 Assesing the quality of visualizations

As proposed in (Sarwate, 2017), quality of the visualization is assessed using 3 main metrics. These are:

- The silhoutte coefficient for the testing of clusters based on the ground truth labels. (*sklearn metrics silhouette score*, n.d.)
- A measure of roudness as presented by Polsby, Popper (Polsby & Popper., 2008).
- The overlap as calculated by the convex hulls for each cluster.

Because the algorithm is expected to run on a local system, the total running time of the pipeline will also be taken into account.

## 1.4 The software package

After selecting algorithms which optimize for the selected metrics, a software package will be developed, enabling the visual exploration of local sound file collections.

# References

Bosch, J. J., Fuhrmann, F., & Herrera, P. (2014, September). *IRMAS: a dataset for instrument recognition in musical audio signals.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.1290750` doi: 10.5281/zenodo.1290750

Fedden, L. (2017). *Comparative audio analysis with wavenet, mfccs, umap, t-sne and pca.* Retrieved from `https://medium.com/@LeonFedden/comparative-audio-analysis-with-wavenet-mfccs-umap-t-sne-and-pca-cb8237bfce2f`

Lostanlen, V., Cella, C.-E., Bittner, R., & Essid, S. (2019, September). *Medley-solos-DB: a cross-collection dataset for musical instrument recognition.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.3464194` doi: 10.5281/zenodo.3464194

McInnes, L., Healy, J., & Melville, J. (2018). *Umap: Uniform manifold approximation and projection for dimension reduction.*

Polsby, D. D., & Popper., R. D. (2008). The third criterion: Compactness as a procedural safeguard against partisan gerrymandering. *Yale Law & policy Review 9.2 (1991): 301-353.*, *9*, 2579-2605.

Sarwate, L. H. H. . A. (2017). *Klustr: A tool for dimensionality reduction and visualization of large audio datasets.* Retrieved from `https://github.com/lamtharnhantrakul/klustr`

*sklearn metrics silhouette score.* (n.d.). Retrieved from `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html`

van der Maaten, L., & Hinton, G. (2008). Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, *9*, 2579-2605.