COMPARISON OF 2-D TIMBRE SPACES FOR FREESOUND SOUNDS

A PREPRINT

Gonzalo Nieto Montero
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
gonzalo.nieto01@estudiant.upf.edu

March 29, 2021

ABSTRACT

The present paper compares four different 2-D timbre spaces to visualize Freesound sounds, result of the combination of two feature extraction methods and two dimensionality reduction techniques. The first feature extraction method uses hand-crafted features, while the second one uses a pretrained VGGish model as feature extractor (AudioSet features). For reducing dimensionality, PCA and t-SNE techniques are used. The evaluation partition of FSD50K, which is ground truth annotated, is employed. Since these annotations are meant for audio event classification and not for timbre classification, only pairs of contrasting and specific labels are selected: Singing-Gunshot_and_gunfire and Fart-Bell. The spaces are evaluated comparing four clustering methods to the ground truth partitions, utilizing the Adjusted Mutual Information score. Results seem to follow the tendency, as suggested by the literature, that AudioSet features help separate timbres better than hand-crafted features. However, further experiments with more label pairs should be performed to be able to draw any significant conclusion.

***K*eywords** FSD50K · Freesound · Timbre spaces · AudioSet · Feature extraction · Dimensionality reduction

1 Introduction

One of the most common usage of the internet nowadays is media sharing, and audio is not an exception. Started as an academic research project in 2005, *Freesound.org* is an online collaborative database of sounds shared by users under Creative Commons licenses [1]. Since 2011, the platform offers an API which allows developers to access its content in novel ways. As of February 2021, more than 40 projects and over 250 research papers are using or have used Freesound data [2]. One good example of this is the Freesound Explorer [3], a web-based visual interface for exploring Freesound in a 2-dimensional space while creating music at the same time. The present paper is in the context of developing a new application for Freesound: a concatenative synthesizer.

The term "concatenative synthesis" has been extensively used to characterize musical systems that produce sound by automatically reusing existing ones, according to some well-defined collection of criteria and algorithmic procedures [4]. It is a common practice to present a 2D representation of sounds as an interface for the user to manipulate the concatenation [5].

How this 2-D space is presented, nonetheless, is not trivial. When working with large libraries, the task of selecting the most suitable sounds for individual user needs can be time consuming and can struggle to retrieve any sounds of interest. Favory et. al. [6] suggest that using embeddings from a neural network model trained on AudioSet [7] might deliver the most appropriate features for clustering sounds, which may be related to a qualitatevely good representation. The goal of this work is to compare these embeddings with hand-crafted features that are know to be relevant in timbre perception, as well as combining them with two dimensionality reduction methods: Principal Components Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE).

Table 1: Selected labels counts in FSD50K.eval

Label	Number of sounds
Singing	194
Gunshot_and_gunfire	134
Fart	97
Bell	396

2 Dataset

Given the Freesound context of the comparison, FSD50K [8] seems like the most appropriate dataset for the present task. FSD50K is an open dataset of human-labeled (ground truth annotated) sound events containing around 50,000 Freesound clips unequally distributed in 200 classes drawn from the AudioSet Ontology.

However, the dimensions of this dataset are excessive for the size of the task, so only the evaluation partition of FSD50K is used. Another reason to choose the evaluation partition is that it is exhaustively labeled, meaning that labels are correct and complete for the considered vocabulary.

It must be noted, however, that the AudioSet ontology is meant for audio event classification, which is a slightly different task than timbre classification. For this reason, not all labels are valid for the present task. As a solution, two pairs of labels are selected following these criteria:

- Labels should be specific enough, so they share a common timbre and are not mixed with other sounds (in the majority of the cases at least).
- Labels should be contrasting enough, so they have 'opposite' timbres.
- Labels should be big enough in number of samples, so they are significant.

The pairs of labels selected consequently are: Singing vs. Gunshot_and_gunfire and Fart vs. Bell. Both cases are pitched and noisy sounds vs. unpitched and tonal sounds. More over, a maximum of 10 seconds duration is set, to increase the chance of obtaining one-shot sounds and because the final application (concatenative synthesis) allows it, as these sounds will be cut in smaller pieces. The final count of sounds for the selected labels is shown in Table 1.

3 Methodology

Two feature extraction methods and two dimensionality reduction methods are compared. The feature extraction methods are a hand-crafted features one, and one that uses what is called in this paper "AudioSet features" (name used from this point on, for the sake of briefness).

The hand-crafted method consists of extracting a set of features that are know to be relevant in timbre perception using Essentia. These selected features are: MFCCs, spectral centroid, spectral spread, spectral complexity, dissonance, log attack time and pitch; which results in a 36-dimensional tensor. For the second feature extraction the VGGish pre-trained Tensorflow model is used as an audio feature extractor. This deep convolutional neural network was trained with the Audioset dataset, that is conformed by 2M YouTube audios for the task of general audio tagging. VGGish takes the log mel spectrogram of an audio as input and converts it into a semantically meaningful, high-level 128-dimensional embedding. Both of these extracted features are averaged across time, to obtain tensor per sound. Lastly, they are normalised, so that the values of different features are in the same range.

Afterwards, the multidimensional features (128-D the AudioSet ones, 36-D the hand-crafted ones) are converted to a 2-dimensional reduction, so that they can be visualized. Here is where PCA and t-SNE techniques are used and compared.

The combination of these methods results in four spaces: (i) hand-crafted features and PCA, (ii) hand-crafted features and t-SNE, (iii) AudioSet features and PCA and (iv) AudioSet features and t-SNE. Furthermore, these four methods are applied to the two pairs of selected labels, resulting in a total of 8 combinations.

All code used for this paper is available at [9].

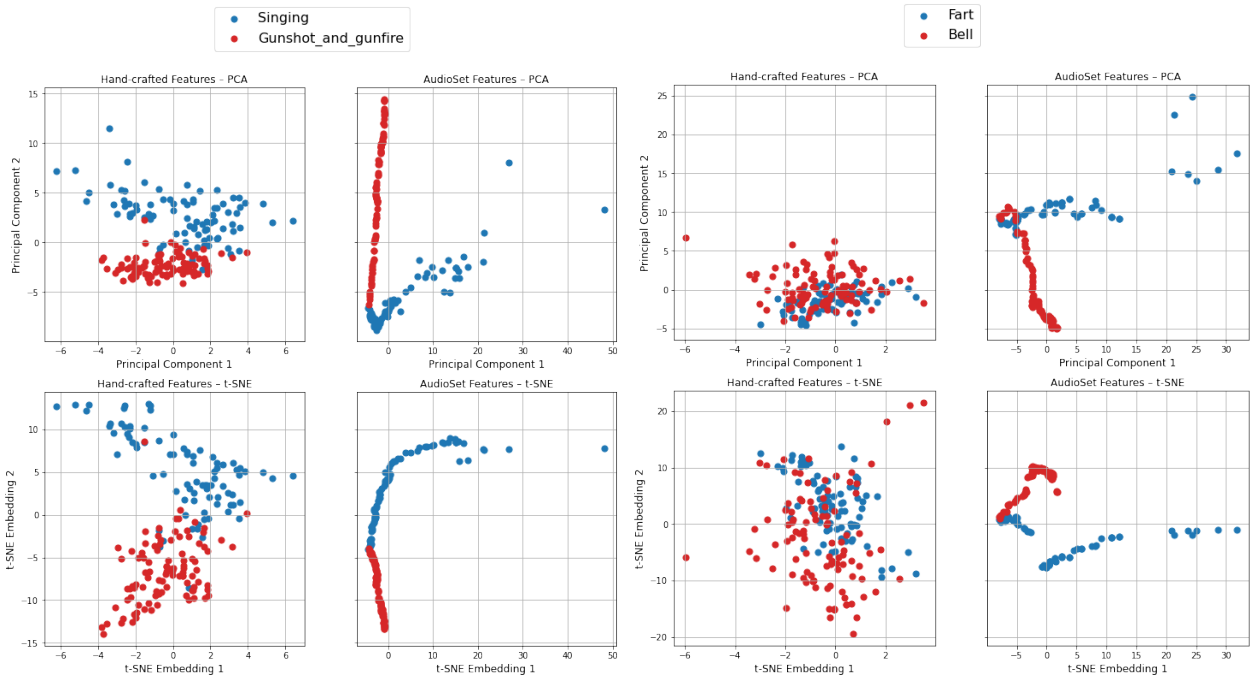

Figure 1: Label pair Singing-Gunshot_and_gunfire (left) and Fart-Bell (right) processed with the four methods. Ground truth labels.

4 Evaluation

For evaluating the different methods used, the processed features are first clustered. Then, the similarity between this clustering and the real partition (given by the ground truth labels) is measured. To measure this similarity, the Mutual Information score adjusted for chance (AMI) is used, which –the literature suggests [8]– is suited when the reference clustering (the ground truth) is unbalanced and there exist small clusters.

The Mutual Information metric (MI) quantifies the information shared by two partitions. When adjusted for chance, the metric takes a value of 1 for two identical partitions and the value of 0 for two randomly dissimilar partitions.

Nonetheless, each clustering method makes some implicit assumptions on the data. These are assumptions that the data may or may not obey. For this reason, several clustering methods are used to evaluate: K-means clustering, spectral clustering, OPTICS and Agglomerative clustering.

5 Results

Once the features are processed following the four methods described in section 3, they are represented with their ground truth labels to have an idea of the distribution of the data. Figure 1 shows these representations.

It can be observed that AudioSet features seem to group the contrasting timbres better than hand-crafted ones, both with PCA and t-SNE reductions. Handcrafted features fail to cluster the label pair Fart-Bell. Between PCA and t-SNE, however, there does not seem to be a considerable difference.

Lastly, all AMI scores are presented in Figure 2. The clustering method clearly conditions the score obtained. K-means and agglomerative clustering methods perform the best, and AudioSet features seem to separate the labels better for the case of Fart vs. Bell.

6 Discussion

It can be observed that the clustering method has a direct influence in the AMI score obtained in the evaluation. This makes sense, as each clustering method makes some assumtions on the data, resulting in successful or unsuccessful clusterings. Therefore, it is important to compare between different clustering methods.

	Agglomerative		K-means		OPTICS		Spectral	
Method	**Singing vs. Gunshot**	**Fart vs. Bell**	**Singing vs. Gunshot**	**Fart vs. Bell**	**Singing vs. Gunshot**	**Fart vs. Bell**	**Singing vs. Gunshot**	**Fart vs. Bell**
Hand-crafted features + PCA	0.61218	0.15641	0.59016	0.16603	0.18646	0.09231	0.25140	0.02497
Hand-crafted features + t-SNE	0.79391	0.06943	0.71947	0.07482	0.22264	0.13973	0.79391	0.00156
AudioSet features + PCA	0.59895	0.72942	0.46759	0.63944	0.28029	0.22796	0.00125	0.01627
AudioSet features + t-SNE	0.70902	0.71829	0.59895	0.60541	0.25809	0.24028	0.70902	0.37386

Figure 2: AMI scores for all extraction methods, both pair of labels and the four clustering methods.

K-means and agglomerative clustering methods seem to deliver similar results in this case (also because they are similar methods), suggesting that they provide correct clusters for the data. Moreover, these results agree with what can be observed in the plotting of ground truths, where the hand-crafted features both with PCA and t-SNE reductions of the label pair Fart-Bell seem difficult to cluster –thus, they score close to zero–; whereas the AudioSet features seem separable –so they score closer to one.

Another observation is that the AudioSet features, obtained with the VGGish pretrained model, seem to distinguish timbre considerably better than hand-crafted features for the case of Fart vs. Bell, as is suggested in the consulted literature. However, the former features perform slighly better than the latter ones in the case of Singing vs. Gunshot_and_gunfire. In any case, it is difficult to know whether this difference is significant or not, and consequently more contrasting labels should be compared as further work.

Regarding the PCA and the t-SNE dimensionality reduction methods, there does not seem to be a significant difference between them –although once again, it is intricate to measure significance having only four scores available. Nonetheless, it has been observed during the implementation of the t-SNE method, that the random initialization of this method varies the output substantially. This means that the clustering carried out after t-SNE could perform better with a different initialization, which is not known a priori. PCA is therefore more stable when compared to t-SNE.

In addition to this, there are other factors that have not been taken into account, such as the interpretability and time efficiency of each method. This could be further work in the comparison of the different 2-D timbre spaces.

References

[1] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*, pages 411–412, 2013.

[2] Freesound Labs. `https://labs.freesound.org/`. Accessed: 2021-02-18.

[3] Frederic Font and Giuseppe Bandiera. Freesound Explorer: Make Music While Discovering Freesound! *Proceedings of the Web Audio Conference*, 2016.

[4] Diemo Schwarz. Concatenative sound synthesis: The early years. *Journal of New Music Research*, 35(1):3–22, 2006.

[5] Cárthach Ó Nuanáin, Sergi Jordà, and Perfecto Herrera. Towards User-Tailored Creative Applications of Concatenative Synthesis in Electronic Dance Music. *International Workshop on Musical Metacreation (MUME)*, 2016.

[6] Xavier Favory, Frederic Font, and Xavier Serra. Search result clustering in collaborative sound collections. *ICMR 2020 - Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 207–214, 2020.

[7] Shawn Hershey, Sourish Chaudhuri, Daniel P.W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 131–135, 2017.

[8] Eduardo Fonseca, Student Member, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K : an Open Dataset of Human-Labeled Sound Events. pages 1–24, 2020.

[9] Gonzalo Nieto Montero. Github repository: Comparison of 2-D timbre spaces for Freesound sounds. `https://github.com/gonznm/2-d-timbre-spaces-for-Freesound`. Accessed: 2021-03-29.