Mapping Timbre Space in Regional Music Collections using Harmonic-Percussive Source Separation (HPSS) Decomposition

Kaustuv Kanti Ganguli[1†], Christos Plachouras[1], Sertan Şentürk[2], Andrew Eisenberg[1] and Carlos Guedes[1]
[1] Music and Sound Cultures research group, New York University Abu Dhabi, UAE
[2] Independent Researcher, UK
[†] Corresponding author: kaustuvkanti@nyu.edu

Introduction

Timbre — tonal qualities that define a particular sound/source — can refer to an instrument class (violin, piano) or quality (bright, rough), often defined comparatively as an attribute that allows us to differentiate sounds of the same pitch, loudness, duration, and spatial location (Grey, 1975). Characterizing musical timbre is essential for tasks such as automatic database indexing, measuring similarities, and for automatic sound recognition (Fourer et al., 2014). Peeters et al. (2011) proposed a large set of audio features descriptors for quantifying timbre, which can be categorized into four broad classes, namely temporal, harmonic, spectral, and perceptual. The paradigms of auditory modeling (Cosi et al., 1994) and acoustic scene analysis (Abeßer et al., 2017; Huzaifah, 2017) also have extensively used timbral features for the classification task. Timbre spaces, in the typical connotation (Bello, 2010), empirically measure the perceived (dis)similarity between sounds and project to a low-dimensional space where dimensions are assigned a semantic interpretation (brightness, temporal variation, synchronicity, etc.). We recreate timbre spaces in the acoustic domain by extracting low-level features with similar interpretations (centroid, spectral flux, attack time, etc.) by employing audio analysis and machine learning.

Based on our previous work (Trochidis et al., 2019), in this paper, we decompose the traditional mel-frequency cepstral coefficients (MFCC) features into harmonic and percussive components, as well as introduce temporal context (De Leon & Martinez, 2012) in the analysis of the timbre spaces. We will discuss the advantages of obtaining the stationary and transient components over the original MFCC features in terms of clustering and visualizations. The rest of the paper is structured in terms of the proposed methodology, experimental results, and finally, the obtained insights.

Method

Ganguli et al. (2020) explored cross-cultural similarities, interactions, and patterns of the music excerpts from the New York University Abu Dhabi Library Collection — the NYUAD music compendium — a growing collection with approximately 3000 recordings from the Arab world and neighboring regions, with a view to understanding the (dis)similarities by employing visualization and dimensionality reduction techniques. This study was limited to unsupervised clustering, and no experiments were carried out on derived temporal features (except log-STFT in entirety). Nevertheless, the timbre space model we applied successfully separated the data into meaningful clusters. A data-driven optimization showed that K=5 clusters (via K-Means clustering) captured the diversity of the corpus without over-fitting. The qualitative evaluation revealed interesting structures. For instance, one cluster included traditional instrumental string music and two other traditional Arab vocal, electronic, and pop music. Folk music excerpts with similar instrumentation from both the two archives were clustered together in the mapping. Figure 1 (left) shows the 2-dimensional t-Stochastic Neighborhood Embedding (t-SNE) representation of the timbre space for K=5, the encircled regions are similar in terms of the above descriptors. Figure 1 (right) shows that the intensity feature has a linear gradient. This trend indicates that there is a systematic variation of the intensity value along one axis which could help in interpreting the other axis by reverse engineering. The computation of intensity depends on the power spectrogram frames and differs upon decomposition of the spectrogram into its stationary and transient components.

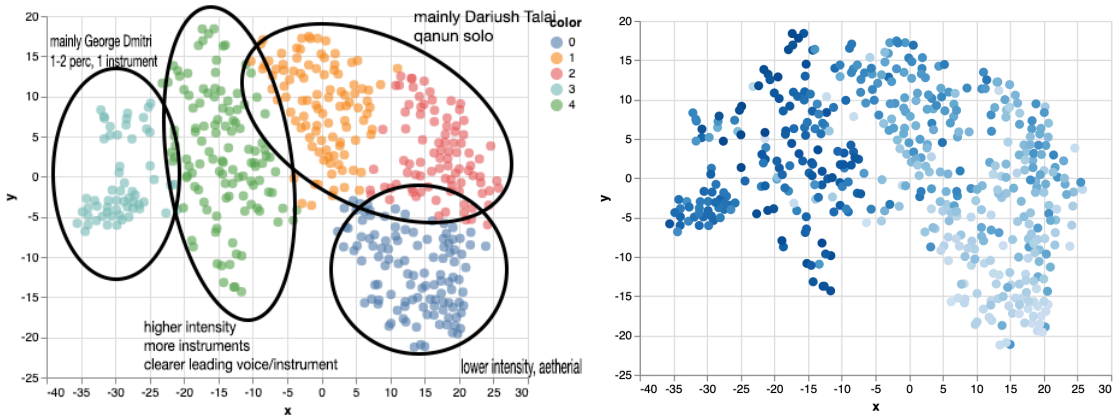

Figure 1: 2D t-SNE representation of the timbre space with K-Means (K=5) clustering on MFCC features, the encircled regions are marked as similar during qualitative evaluation (left). The intensity feature (computed as time-average of frame-wise energy sum) shows a linear gradient (right).

Fitzgerald (2010) proposed a harmonic-percussive source separation (HPSS) method commonly used in music information retrieval (MIR) to suppress transients when analyzing pitch content or suppress stationary signals when detecting onsets or other rhythmic elements. As the music corpus under study comprises a balanced mixture of harmonic and percussive instruments, we employ HPSS[1] to obtain two power spectrograms from each audio excerpt. Figure 2 (left) shows the power spectrogram and the derived harmonic/percussive components for a case-study excerpt. The corresponding MFCC vectors (n_mfcc=13) and the delta feature for the percussive feature vector are also shown in Figure 2 (right).

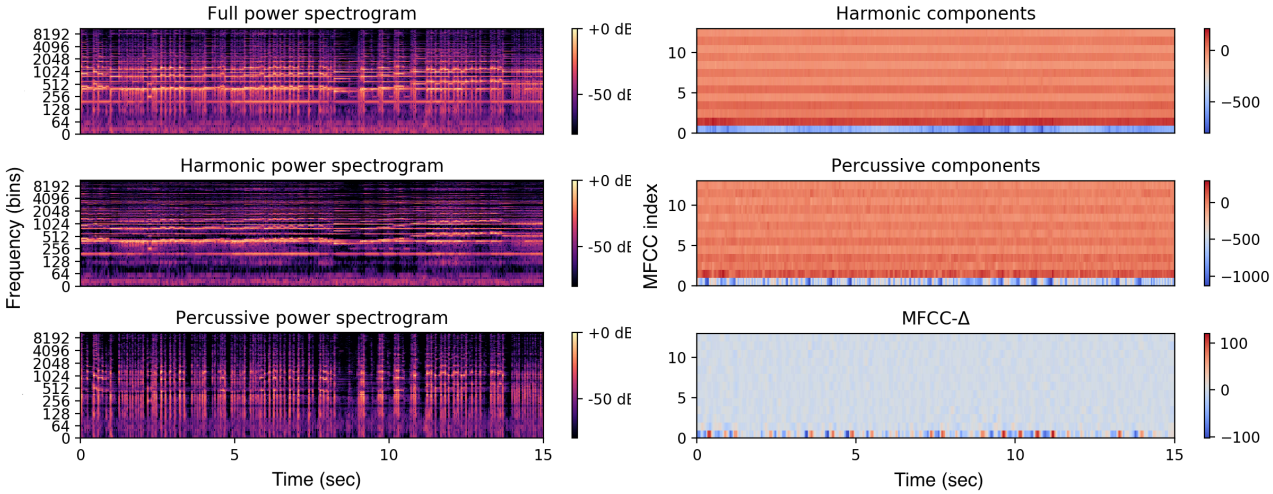

Figure 2: The power spectrogram and the derived harmonic/percussive components for a case-study excerpt (left). The MFCC vectors and the delta feature for the percussive feature vector (right).

The given case-study excerpt consists of both vocal and both melodic and percussive instruments. Some of the melodic instruments are also plucked-string in nature which raises an interesting scenario for the analysis — these instruments produce a wide-band attack at the onset and a stationary sustain before a fading release. The bass drums of the percussive instruments are not pronounced; we can thus safely assume these to be purely transient signals. This phenomenon is visible in Figure 2 (left) in the percussive power spectrogram where the transients are caused by both percussion and plucked-string melodic instruments.

[1] Using LibROSA (McFee et al. (2015)) version 0.8.0, DOI; 10.5281/zenodo.3955228 :: librosa.decompose.hpss

Results

We evaluate the clustering performance in terms of a homogeneity metric for the 2D t-SNE rendering for different spectrogram components. The cluster purity, which is defined as a measure of the extent to which clusters contain a single homogeneous class (Manning et al., 2008), is indicative of the timbre space being able to capture the intended groupings. The delta (differential coefficients, denoted as Δ) and delta-delta (acceleration coefficients, denoted as ΔΔ) features capture the spectrogram dynamics which is a proxy for a pseudo-temporal evolution of the (speech-like) music signal. Table 1 shows that adding delta features improve clustering performance. The harmonically enhanced features have shown overall better performance compared to its percussive counterpart.

Table 1: Cluster Purity for 2D t-SNE rendering for different components of the spectrogram.

No. of clusters (k)	**Full**		**Decomposed**	
	W/o delta	**W delta**	**Harmonic**	**Percussive**
2	*.84*	*.85*	*.85*	*.81*
4	*.86*	*.88*	*.87*	*.84*
6	*.89*	*.92*	*.91*	*.88*

This is, to some extent, intuitive as timbre modeling broadly captures the spectral envelope. However, there is further scope to experiment on hyperparameter tuning to investigate the percussive components. The double-delta parameters did not show significant differences. Table 1 also shows that the stationary components (harmonic MFCC) yield similar performance compared to the full spectrograms with delta features. It is, however, difficult to infer individual phenomena from the cluster purity metric, which involves a series of transformations and a machine learning model. On a corpus level, our proposed methods show better performance.

Discussion

We reported a modified timbre space for the NYUAD music compendium, a collection of recordings from the Arab world and neighboring regions, where a harmonic-percussive decomposition along with delta MFCC features show improvement in the clustering performance. The harmonic components outperformed the percussive components for the given metric; however, the percussive power spectrogram leads to a better tempo estimation and serves as a more reliable feature in rhythm-based studies. As mentioned before, the plucked-string melodic instruments produce both stationary and transient components, which may be utilized as complementary features. This is particularly important because there is hardly any timbre model found for non-Eurogenetic music cultures, whereas it might be easy to obtain a template for Western instruments. Hence, the proposed framework can be beneficial for regional music collections involving folk instruments. Extending from our previous work (Ganguli et al., 2020), we also plan to regenerate the timbre space model in the VR (virtual reality) space with the 3D t-SNE realization of the harmonic and percussive components obtained from the HPSS. This will lead to two independent timbre spaces that can aid in pedagogical as well as community engagement applications.

One drawback of the proposed approach is that it can only serve as a high-level exploratory data analysis tool since there are still not enough metadata regarding the style, genre, instrumentation, and structure of the archives. However, the HPSS decomposition can be particularly useful for an instrument recognition study. This can be better augmented with an audio thumbnailing task by discovering the predominant lead instrument of an excerpt followed by template matching of the representative frames. Finally, analysis of the perceptual timbre space (McAdams et al., 1995) is proposed as future work.

Acknowledgments

This research is part of the project "Computationally engaged approaches to rhythm and musical heritage: Generation, analysis, and performance practice," funded through a grant from the Research Enhancement Fund at the New York University Abu Dhabi.

References

Abeßer, J., Mimilakis, S. I., Gräfe, R., Lukashevich, H., & Fraunhofer, I. D. M. T. (2017). Acoustic scene classification by combining autoencoder-based dimensionality reduction and convolutional neural networks. In Proceedings of IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2017).

Bello, J. P. (2010). Low-level features and timbre. Lecture notes MPATE-GE 2623 Music Information Retrieval, New York University.

Cosi, P., De Poli, G., & Lauzzana, G. (1994). Auditory modelling and self-organizing neural networks for timbre classification. Journal of New Music Research, 23(1), 71-98.

De Leon, F., & Martinez, K. (2012). Enhancing timbre model using MFCC and its time derivatives for music similarity estimation. In 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 2005-2009, IEEE.

Fitzgerald, D. (2010, September). Harmonic/percussive separation using median filtering. In Proceedings of Digital Audio Effects (DAFX), Vol. 10, No. 4.

Fourer, D., Rouas, J. L., Hanna, P., & Robine, M. (2014). Automatic timbre classification of ethnomusicological audio recordings. In Proceedings of the International Society for Music Information Retrieval (ISMIR) conference, Taipei, Taiwan.

Ganguli, K. K., Gomez, O., Kuzmenko, L., & Guedes, C. (2020). Developing immersive VR experience for visualizing cross-cultural relationships in music. In Proceedings of 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (pp. 401-406). IEEE.

Grey, J. M. (1975). An exploration of musical timbre. Ph. D dissertation, Stanford University.

Huzaifah, M. (2017). Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. arXiv preprint arXiv:1706.07156..

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Evaluation in information retrieval. Introduction to information retrieval, 1, 188-210.

McAdams, S., Winsberg, S., Donnadieu, S., Soete, G. D., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. Psychological Research, 58, 177–192.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). LibROSA: Audio and music signal analysis in python. In Proceedings of the 14th Python in science conference (Vol. 8, pp. 18-25).

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The timbre toolbox: Extracting audio descriptors from musical signals. The Journal of the Acoustical Society of America, 130(5), 2902-2916.

Trochidis, K., Russell, B., Eisenberg, A., Ganguli, K. K., Gomez, O., Plachouras, C., Guedes, C., & Danielson, V. (2019). Mapping the Sounds of the Swahili coast and the Arab Mashriq: Music research at the intersection of computational analysis and cultural heritage preservation, In Proceedings of the Digital Libraries for Music (DLfM) conference, The Hague, The Netherlands.