Exploring Timbre Spaces through Dimensionality Reduction and Clustering

Pablo R. Sene[1], André E. Lazzaretti[1], Heitor S. Lopes[1], Thiago H. Silva[1,2], Bruno S. Chang[1]
[1]Graduate Program in Electrical Engineering and Industrial Informatics (CPGEI), UTFPR
Curitiba, PR, Brazil
[2]University of Toronto, Toronto, Canada
Emails: psene@alunos.utfpr.edu.br, lazzaretti@utfpr.edu.br, hslopes@utfpr.edu.br, thiagohsilva@utfpr.edu.br, bschang@utfpr.edu.br

***Abstract*—This work investigates the relationship between perceptual and computational timbre spaces by applying dimensionality reduction techniques to audio features extracted from musical recordings. A subset of three stylistically related genres—Samba, Jazz, and Afrobeat—was selected, and a comprehensive set of features, including harmonic and percussive descriptors, was extracted. Dimensionality reduction methods such as PCA, t-SNE, and UMAP were used to visualize the resulting timbre space. Clustering algorithms (K-Means and HDBSCAN) were applied to evaluate the consistency of the computed space with genre and cultural groupings. Interestingly, instances of perceptual similarity—such as João Gilberto's 'Desafinado' aligning with Jazz clusters—suggest that the computational space is capable of capturing musically meaningful relationships. The results indicate that appropriate feature design and dimensionality reduction techniques can, to some extent, reflect human-perceived timbral similarities across styles and cultural contexts.**

***Index Terms*—Timbre spaces, Dimensionality reduction, Clustering, Musical analysis, Data visualization**

I. Introduction

Timbre is the perceptual attribute that distinguishes sounds of identical pitch, loudness, and duration, defined by the American National Standards Institute [1] as the quality that differentiates tones from distinct sources. Despite its importance, timbre remains one of the most complex auditory attributes to define and quantify.

Early studies in psychoacoustics and cognitive psychology [2], [3] have shown that timbre perception is multidimensional, involving descriptors such as brightness, warmth, and roughness. These perceptual dimensions were further explored using listener surveys and verbal reports [4], [5], revealing consistent patterns in how people describe and group different timbres.

In parallel, computational approaches have emerged to represent timbre by extracting features from audio signals, aiming to build "timbre spaces" using dimensionality reduction and clustering algorithms [6], [7]. However, the correspondence between these computational representations and perceptual similarity remains an open question.

This study investigates the relationship between perceptual and computational timbre spaces by focusing on three stylistically and rhythmically related musical genres: Samba, Jazz, and Afrobeat. These genres, although culturally distinct, share instrumental and structural characteristics that may produce overlapping timbral features.

By extracting harmonic and percussive descriptors from audio recordings and applying dimensionality reduction techniques (PCA, t-SNE, UMAP), followed by clustering algorithms (K-Means, HDBSCAN), we evaluate how well the computational timbre space aligns with genre distinctions and perceptual similarities. In particular, we explore whether such spaces can capture musical relationships that go beyond simple stylistic labels.

A. Related Work

The study of musical timbre has historically drawn from both perceptual and acoustic perspectives. Foundational works such as [2] and [3] used multidimensional scaling and verbal descriptors to map how listeners perceive timbral differences, establishing the notion of a "timbre space" based on human judgment. Further studies (e.g., [4], [6] contributed to understanding the cognitive and subjective dimensions of timbre, emphasizing its multidimensional nature.

On the computational side, authors such as [7] and [8] explored how spectral and temporal features can be used to characterize timbre in measurable terms. These works highlight the technical challenge of bridging physical signal descriptors with the richness of perceptual experience.

More recent efforts have aimed to combine these two views. [9], for example, mapped the timbral landscape of Arabic music by using Harmonic-Percussive Source Separation (HPSS) and a large set of audio features. Dimensionality reduction techniques were applied to visualize stylistic similarities and cultural relationships, indicating that computational models can reflect aspects of timbral perception.

However, few studies attempt a direct comparison between perceptual expectations and the structures of computationally derived timbre spaces. Additionally, the use of clustering metrics as a means to evaluate the coherence of these spaces—particularly in terms of genre or cultural grouping—remains underexplored. This work seeks to contribute to that gap by combining perceptually informed genre selection with quantitative embedding analysis and unsupervised clustering.

II. Methods

This section describes the methodology adopted to construct and analyze the computational timbre space. The process includes dataset construction, feature extraction, dimensionality reduction, and cluster evaluation. Each step is designed to investigate whether perceptually similar musical pieces are also grouped closely in the computed space.

A. Dataset

In this project, both quantitative and qualitative evaluations were considered. To enable a more perceptual analysis, a limited and well-curated set of songs was selected. The underlying assumption is that meaningful evaluation requires familiarity with the material—knowing the songs and recognizing which ones sound perceptually similar. For this reason, Afrobeat, Jazz, and Samba were selected as the target genres, as shown in table I. Although it may be difficult to articulate precisely what makes them similar, these genres are often perceived as *timbrally* related. This subjective perception provides an interesting scenario to explore the correspondence between perceptual and computational timbre spaces.

Each track was manually tagged with its corresponding musical style and country of origin. Since each style is uniquely associated with a specific country in this dataset (e.g., Jazz from the USA, Samba from Brazil, Afrobeat from Nigeria), these two labels are treated as interchangeable for the purposes of evaluation.

All files were converted to mono WAV format and trimmed to a fixed duration of 200 seconds to reduce variability due to track length.

B. Feature Extraction

To extract the relevant temporal and spectral features of each audio recording, the *Librosa Library* was used. The audio duration was fixed to 200 seconds and resampled to 22,050 Hz to preserve audio fidelity and capture frequencies lower than 11 kHz, which is appropriate for musical analysis. A technique called Harmonic-Percussive Source Separation (HPSS) was used to enable the extraction of specialized features from each component.

Global features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroids, spectral roll-off, and spectral bandwidth were computed directly from the full audio signal. From the harmonic component, chroma and tonnetz features were extracted to capture pitch content and tonal relationships. From the percussive component, zero-crossing rate (ZCR), spectral flatness, and root mean square (RMS) energy were computed to capture rhythmic and noise-like properties.

Mean and standard deviation were computed for each feature, forming one vector with 74 features per song.

Table II presents the features and their corresponding descriptions.

C. Feature Normalization

Before applying dimensionality reduction and clustering algorithms, all audio features were normalized using z-score normalization. This process transforms each feature to have a zero mean and unit variance, preventing features with larger numerical ranges from dominating the results. The normalization was performed using the 'StandardScaler' method from the scikit-learn library.

D. Dimensionality Reduction

To investigate the timbre space visually, three dimensionality reduction techniques were applied to the normalized features: Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP). These techniques transform the 74-dimensional feature vectors into two dimensions, allowing for a representation that captures the most relevant similarities between songs. The default parameters were used in the experiment.

While PCA provides a linear projection based on variance maximization, t-SNE and UMAP offer nonlinear mappings that better preserve local relationships and clustering tendencies within the data.

Regarding the choice of hyperparameters, PCA was applied with its standard configuration, as it has no tunable parameters in two dimensions. For t-SNE, a perplexity of 5 was selected due to the small dataset size, emphasizing local neighborhood preservation. For UMAP, the random seed was fixed to 42 to ensure reproducibility, while other parameters were kept at their default values, which are commonly recommended for exploratory visualization. Preliminary tests with alternative parameter values yielded similar qualitative patterns, indicating that the results are robust to moderate parameter changes.

The parameters used for each technique are presented in Table III.

E. Clustering and Evaluation

After projecting the songs into a two-dimensional space, clustering algorithms were applied to investigate the grouping behavior within the timbre space. Two unsupervised clustering techniques were used: K-Means and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise).

K-Means was configured to form three clusters, reflecting the number of genres in the dataset. This method assigns each point to one of the clusters by minimizing the distance to the cluster centroids.

HDBSCAN, on the other hand, does not require the number of clusters to be defined in advance. Instead, it identifies clusters of varying densities and treats less consistent points as outliers. This flexibility makes HDBSCAN particularly appropriate for exploring the structure

TABLE I
Manually Created Dataset

Style	Songs
Afrobeat	30
Jazz	30
Samba	30

TABLE II
Extracted Audio Features

Feature	Type	Description
MFCC1–13	Global	Mel Frequency Cepstral Coefficients; capture timbral envelope
Spectral Centroid	Global	Indicates the "brightness" of the sound
Spectral Rolloff	Global	Frequency below which a percentage of total energy lies
Spectral Bandwidth	Global	Width of the frequency band in the spectrum
Chroma 1–12	Harmonic	Energy distribution across pitch classes (C to B)
Tonnetz 1–6	Harmonic	Tonal relations in pitch space (tonal centroid features)
ZCR	Percussive	Zero Crossing Rate; noisiness/rhythm
Flatness	Percussive	Spectral flatness; tonality vs noise
RMS	Percussive	Root Mean Square energy; signal amplitude

of the timbre space, especially in the presence of overlapping genre characteristics.

Both clustering methods were applied to the 2D representations produced by PCA, t-SNE, and UMAP. The quality of the resulting clusters was evaluated using internal metrics such as the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index.

Table IV presents the parameters used for K-Means and HDBSCAN.

III. Results and Discussion

The results are divided into two main parts: a quantitative analysis and a qualitative analysis, which are explored in the following subsections. The goal is to validate whether numerical results can reflect human musical perception—that is, whether the computational space aligns with the perceptual space.

A. Overview of the Projections

To explore the structure of the timbre space, the 74-dimensional feature vectors were projected into two-dimensional space using PCA, t-SNE, and UMAP. Each projection provided a distinct view of the data distribution, allowing for both visual and numerical assessment of groupings.

The PCA projection showed limited separation between styles, forming a single dense cluster with no clear boundaries. This was expected, as PCA is a linear method and may not effectively preserve local relationships in complex, perceptually grounded data like timbre.

TABLE III
Hyperparameters used for dimensionality reduction

Method	Parameters
PCA	—
t-SNE	perplexity = 5, random_state = 42
UMAP	random_state = 42

TABLE IV
Hyperparameters used for Clustering

Method	Parameters
K-Means	n_clusters = 3
HDBSCAN	min_cluster_size = 8

In contrast, t-SNE revealed more localized groupings. While some overlap between styles occurred—which is expected, given that the dataset was intentionally composed of perceptually similar genres—small clusters appeared, particularly around Jazz recordings.

The UMAP projection provided the most visually consistent structure, with two better-defined clusters and one more diffuse group, as shown in Figure 3. One cluster was clearly associated with Jazz recordings, aligning with perceptual expectations.

To quantitatively evaluate the effectiveness of each projection, three unsupervised clustering metrics were computed using the known style labels: Silhouette Score ([10]), Calinski-Harabasz Index ([11]), and Davies-Bouldin Index ([12]). These indices are widely adopted as internal validation measures in clustering tasks. As shown in V, UMAP achieved the best overall performance across all three metrics, reinforcing the visual impression of a better-defined stylistic structure.

B. Clustering

As discussed in the previous sections, two clustering techniques were applied to the low-dimensional representations of the musical samples—K-means and HDBSCAN. Each projection and clustering method combination was evaluated using the same validation metrics as before: Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. In this context, these metrics provide insight into the spatial arrangement and quality of the resulting clusters.

However, it is important to note that no evaluation metrics were reported for the combinations of HDBSCAN with PCA or t-SNE. In these cases, HDBSCAN was unable to form valid clusters, assigning all or nearly all data points to noise (i.e., the label -1). Since the clustering metrics rely on the existence of multiple distinct groups, the absence of valid clusters makes quantitative evaluation meaningless for those configurations.

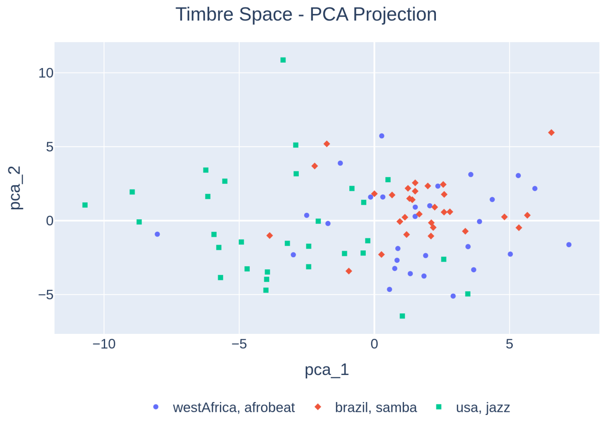

Fig. 1. Timbre Space – PCA

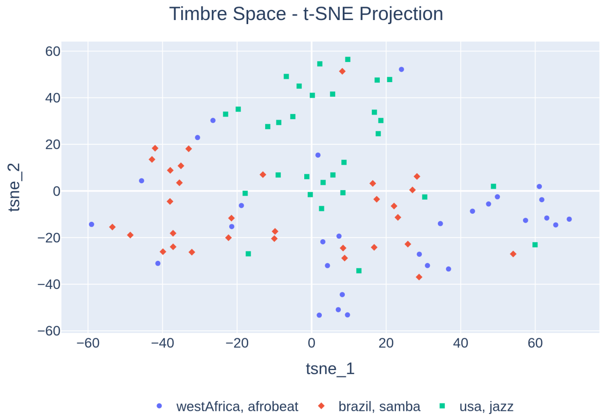

Fig. 2. Timbre Space – t-SNE

Among the clustering results (Table VI), UMAP combined with HDBSCAN (Figure 5) presented the most coherent structure, achieving the highest Silhouette Score (0.535), the highest Calinski-Harabasz Index (55.737), and the lowest Davies-Bouldin Index (0.529). These values suggest compact, well-separated clusters consistent with the data's intrinsic structure.

In contrast, PCA followed by K-Means yielded the weakest results, particularly a Davies-Bouldin Index of 29.412, indicating poorly defined clusters with high overlap. This aligns with previous observations that PCA fails to expose meaningful separation in the data.

The t-SNE + K-Means and UMAP + K-Means combinations showed intermediate performance, with UMAP again outperforming t-SNE in all metrics. This supports the idea that UMAP is more effective at preserving both local and global structures in the reduced space.

C. Qualitative Analysis

Besides the numerical results, listening to the songs helped to confirm that the computational timbre space captured meaningful similarities between some tracks.

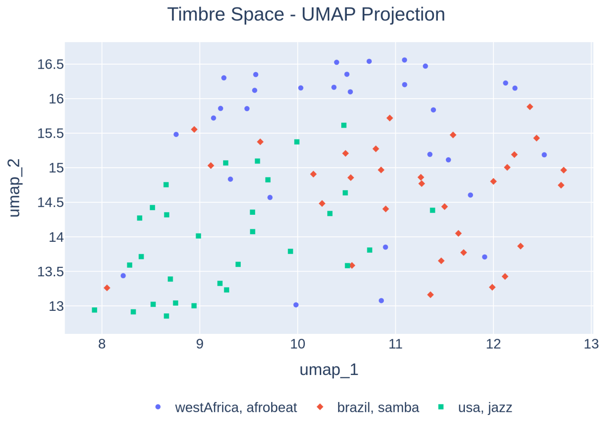

Fig. 3. Timbre Space – UMAP

TABLE V
QUANTITATIVE EVALUATION OF DIMENSIONALITY REDUCTION TECHNIQUES

Method	Silhouette	Calinski–Harabasz	Davies–Bouldin
PCA	0.050	16.537	3.429
t-SNE	0.064	9.440	2.715
UMAP	0.138	22.736	2.085

One interesting case is the song "Desafinado". The version recorded by João Gilberto with the American saxophonist Stan Getz was grouped with Jazz songs—even though "Desafinado" is originally a Brazilian samba. Listening to this version sounds closer to Jazz: the harmony, the smooth rhythm, and the saxophone give it a different character. The model reflected that, placing it next to other Jazz songs.

On the other hand, another version of "Desafinado", this time instrumental by Stan Getz and Charlie Byrd, was placed in a different cluster. This version sounds more like a traditional samba, with more rhythmic elements and a stronger percussive feel. Curiously, the version recorded

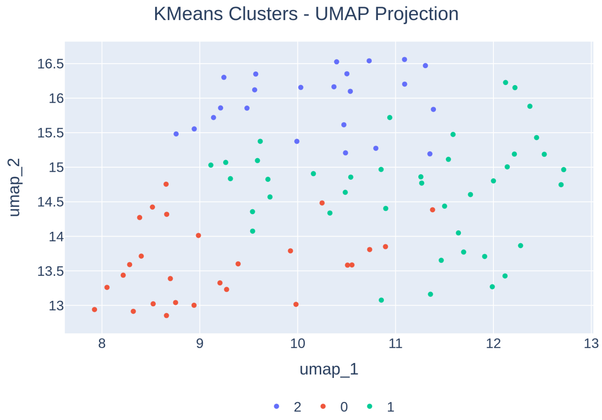

Fig. 4. K-Means and UMAP Clusters

TABLE VI
Clustering Evaluation Metrics

Method	Silhouette	C-H	D-B
PCA + K-Means	0.095	31.722	29.412
t-SNE + K-Means	0.195	33.381	1.393
UMAP + K-Means	0.260	46.761	1.195
UMAP + HDBSCAN	0.535	55.737	0.529

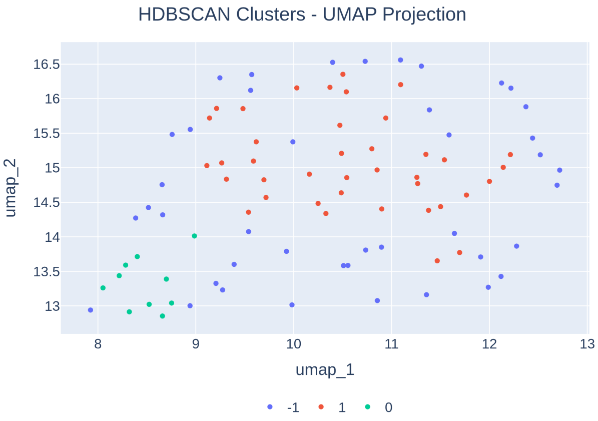

Fig. 5. HDBSCAN and UMAP Clusters

by a Brazilian singer sounds more like Jazz, while the one recorded by two Americans sounds more like Samba— but this shows how the performance can change the perceived style.

Other Jazz songs were also grouped together, especially the ones with similar instruments and soft textures. In contrast, some Samba and Afrobeat tracks were spread out, possibly because these styles have more variety, or the features used were not enough to capture their essence.

These examples suggest that even with a small dataset and simple features, the timbre space was able to capture some real musical similarities that make sense to the human ear.

IV. Conclusion

This project investigated how computational models can represent the way humans perceive timbre. By applying dimensionality reduction and clustering techniques to a selected group of songs, visual representations of timbre space could be constructed. Among the tested methods, nonlinear techniques like t-SNE and UMAP showed better results than PCA, especially when forming clear and meaningful groups.

The best cluster coherence was achieved using UMAP together with HDBSCAN, which produced compact and stylistically consistent clusters. One interesting example is how João Gilberto's "Desafinado" appeared alongside American Jazz tracks, showing that the model captured subtle similarities between genres from different cultures.

Although the results are promising, it is important to note some methodological limitations of this study. The dataset was intentionally small and composed of perceptually similar genres, which enabled controlled analysis but limited statistical generalization. Moreover, since each musical style corresponds to a specific country (e.g., Samba–Brazil, Jazz–USA, Afrobeat–West Africa), the style and country labels are effectively interchangeable, potentially introducing cultural bias.

Despite these limitations, the approach presented here shows potential for a better understanding of musical timbre and could support future work in music analysis, cultural studies, and recommendation systems.

Acknowledgments

This research was partially supported by CNPq (processes 314603/2023-9, 441444/2023-7, 409669/2024-5, 444724/2024-9, and 405940/2022-0).

References

[1] American National Standards Institute, *American National Standard: Psychoacoustical Terminology.* New York: ANSI, 1973.

[2] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277, 1977.

[3] G. Von Bismarck, "Timbre of steady sounds: A factorial investigation of its verbal attributes," *Acustica*, vol. 30, no. 3, pp. 146–159, 1974.

[4] R. L. Pratt and P. E. Doak, "A subjective rating scale for timbre," *Journal of Sound and Vibration*, vol. 45, 1976.

[5] A. Zacharakis and J. Reiss, "An exploratory study of musical timbre through verbal descriptions," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2011. [Online]. Available: https://www.eecs.qmul.ac.uk/~josh/documents/2011/ZacharakisReiss-2011-ISMIR.pdf

[6] S. McAdams and E. Bigand, *Thinking in Sound: The Cognitive Psychology of Human Audition.* Oxford: Oxford University Press, 1993.

[7] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale.* New York: Springer, 2005.

[8] D. M. Howard and J. A. S. Angus, *Acoustics and Psychoacoustics*, 5th ed. New York: Routledge, 2017.

[9] C. Guedes, K. Ganguli, C. Plachouras, S. Senturk, and A. Eisenberg, "Mapping timbre space in regional music collections using harmonic-percussive source separation (hpss) decomposition," in *Proceedings of the 2nd International Conference on TImbre*, Sep. 2020.

[10] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[11] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[12] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.