DOI: 10.61782/fa.2023.0501

forum acusticum 2023

MUSICAL TIMBRE CLASSIFICATION USING FFT-ACOUSTIC DESCRIPTORS AND MACHINE LEARNING

Yubiry González* **Ronaldo Prati**
Center of Mathematics, Computer Science and Cognition at Federal University of ABC, Av. Dos Estados 5001, Santo André 09210-580, SP, Brazil

ABSTRACT

Musical timbre is a complex multidimensional attribute of auditory perception, which allows, in a first approximation, to discriminate between musical instruments when they have the same sound, intensity, and duration. Also, in some cases, there are sounds that appear to have very close timbral similarity, even when the instruments have different acoustic characteristics. This fact can make it difficult to classify musical instruments by timbres. We explore a 7-dimensional abstract space, formed by the fundamental frequency and acoustic descriptors extracted from Fourier Transform in five musical instruments: Trumpet, violin, cello, transverse flute, and clarinet, of a monophonic audio record, from the Tinysol and Good-Sounds databases, corresponding to the fourth octave. This approach makes it possible to define a collection of points in timbral space uniquely and allows differentiating sounds played in ordinary style on any type of musical instrument. Through the geometric distance between musical sounds, we explore some Machine Learning techniques to establish categories of similarities between musical sounds, instruments, and family of musical instruments. It is concluded that the study of timbral similarity through geometric distances made it possible to find clustering between categories of musical timbre.

Keywords: *Musical timbre; FFT; musical instruments; acoustic descriptors; Machine Learning; Data Analysis; TinySol; GoodSounds.*

* ***Corresponding author***: *yubiry.ggonzalez@gmail.com*

1. GENERAL OVERVIEW: FFT-ACOUSTIC DESCRIPTORS

Musical timbre is a generic multidimensional attribute that allows distinguishing between sounds of the same pitch, intensity, and duration. Timbre characterization is important for the identification and classification of musical instruments, for audio analysis and synthesis, and in general for Computer-Assisted Composition [1-2]. The study of musical timbre can be done from various perspectives, both from acoustics and psychophysics [3-4]. Analysis of audio recordings can also be used, both in the domain of time (spectrograms) and in frequencies (FFT). However, the digitization of sounds is based on the Fourier transform of audio recordings, where the relevant aspects of musical timbre are necessarily contained in the collection of amplitude and frequency pairs that emerges from the FFT of monophonic audios, with the independence of the descriptions that can be made from the domain of time (spectrograms) and of the psychoacoustic approach to hearing.

On the other hand, in Western orchestral music, the musical sounds form a succession of finite and well-defined frequencies (tempered scale). Therefore, the audio recording of each musical note and musical instrument, with a specific dynamic (pianissimo, mezzoforte, fortissimo) provides a unique FFT that characterizes it. This FFT will be univocally prescribed by the fundamental frequency (f_0) and a specific collection of pairs of numbers that correspond to the frequencies and amplitudes of the partial components (harmonic or not) coming from the spectral decomposition of the audio signal. Therefore, the musical timbre information must be contained in this set of frequencies and amplitudes associated with the digital recording of the audio.

2. METHODOLOGY

The open-source sound library Good-Sounds® [9] and Tinysol® [10] provide monophonic audio recordings of real instruments, played by professional musicians. From these libraries, 256 recordings were selected in the WAV audio format, corresponding to the instruments Violin, Cello, Transverse Flute, Clarinet, and Trumpet, in the fourth octave of the equal temperament scale and in the mezzoforte dynamic, played in the so-called "ordinary" style and in the absence of a mute. For each recording, the FFT is obtained with normalized amplitudes, using the SciPy library module in Python [11]. Timbral coefficients are calculated from the FFTs, which are dimensionless, univocal, and independent descriptors of the FFTs [6,8]. These six timbral coefficients, together with the fundamental frequency (f_0) provide, for each audio record, a seven-dimensional vector that defines a point in an abstract space, which also is a geometric space. In this timbral space, points close to each have similar coordinates and, therefore, similar timbral coefficients, and consequently similar timbral properties, for details see Figure 1.

Musical sounds with similar fundamental frequency (f_0) correspond to analogous sounds in the tempered scale. For a specific musical instrument, the relative measure of the amplitude of the fundamental frequency with respect to the set of amplitudes of the FFT (Affinity Coefficient A) and the average variation of the envelope of the pulses in the FFT (Monotonicity coefficient M) are associated to the musical octave [8], the difference in the composition of harmonics (Spectral Signature) and the average value of the harmonicity of the partial frequencies (Harmonicity coefficient H) allow to identify the musical instrument [6]. For a given musical instrument and a specific musical sound, the relative measure of the amplitude of the fundamental frequencies (Sharpness coefficient S, note that this is not Zwicker's psychoacoustic sharpness) and the average of the deviation of the amplitudes of the partial frequencies with respect to the amplitude of the fundamental (MA Coefficient) report dynamics [8].

However, different musical sounds played by different instruments can be perceived as timbrically similar, and therefore should be close in timbral space [7].

To find these similarities in the timbral representation, the data set was partitioned by types of instruments, families, and musical notes, in order to group them and obtain their characteristic mean values. The K-means algorithm was applied to these data, which uses Euclidean distance as a metric and variance as a measure of group dispersion. The K-means algorithm iteratively clusters data points by minimizing the sum of squares within the cluster, thus being a simple and effective clustering method. Although the K-means algorithm tends to generate clusters of similar size and spherical shape, we did not find large differences when using general techniques such as the Gaussian Mixture Model Algorithm (GMM). The results of the exploratory analysis with K-means to distinguish groups of timbral similarities are shown in the next section.

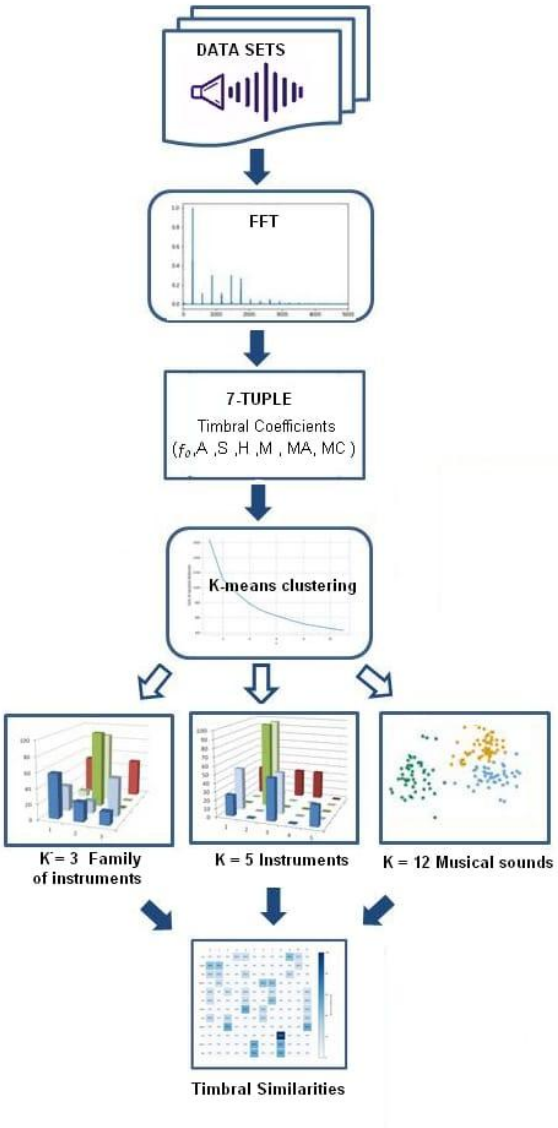

Figure 1. General procedure for timbral similarities using FFT and Machine Learning.

3. RESULTS AND DISCUSSION

Three categories or partitions of the data set were studied: (3.1) five clusters for the five Musical Instruments, (3.2) three clusters to delimit the three families of Instruments: Wooden Aerophones (Flute and Clarinet), Metal Aerophones (Trumpet), and String Instruments (Violin and Cello), and (3.3) twelve clusters for the musical sounds of the fourth octave, in each instrument separately and globally for the set of instruments. The mean values of the 7-dimensional tuples (fundamental frequency and timbral coefficients) were computed for the audio recordings with the same instrument and musical sound.

3.1 Clustering by Instruments

Figure 2 shows the percentage distribution of data in each cluster for the various musical instruments under consideration. Even though the data is not grouped into a single cluster for each musical instrument, we can observe that the data for each musical instrument appear in very specific, bounded regions of the timbral space. For example, all the sounds of the Flute and the Clarinet appear in clusters 4 and 5, respectively, while those of the trumpet only appear in clusters 2 and 3. None of the data for the Violin and Cello are found in cluster 5. The value of the inertia (a measure of how internally coherent the data points are within each cluster) of this distribution is 147.83, which is equivalent to saying that the average distance between a datum and the center of the cluster is 12.2, much lower than the distances between clusters 27.5, 109.3, 39.2, and 125.21 for the centers of cluster 1, 2 4 and 5 respectively. A detailed analysis of each timbral subspace by the instrument can be performed (omitted for brevity of the report) from these values by considering the coordinates of each centroid and its mean variance of 12.2.

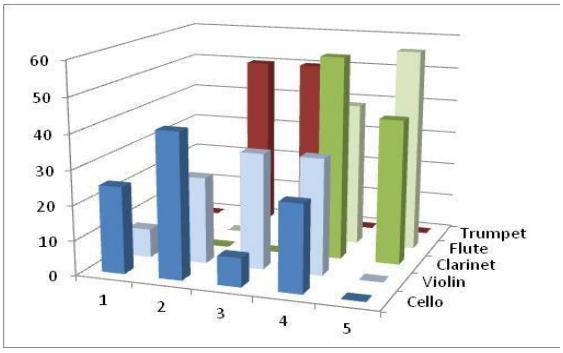

Figure 2. K-Mean clustering by musical instruments in percentages of data by instrument and for both databases.

The same partitioning of the data set was performed for the Tinysol and Good-Sounds databases separately, and results are shown in Figure 3. The results show similar clustering, although the absolute distributions vary, as do the centroids of each cluster, as well. It is observed in both bases that the Flute and the Clarinet occupy only clusters 4 and 5; that none of the data for the violin, the cello, and the trumpet appear in cluster 5, and that the majority of the trumpet occupies only two clusters (Note that the centroids of each cluster vary in both plots of Figure 3). The inertia values are 144.9 and 130.3, which reports that the dispersion is smaller in the Good-Sounds database.

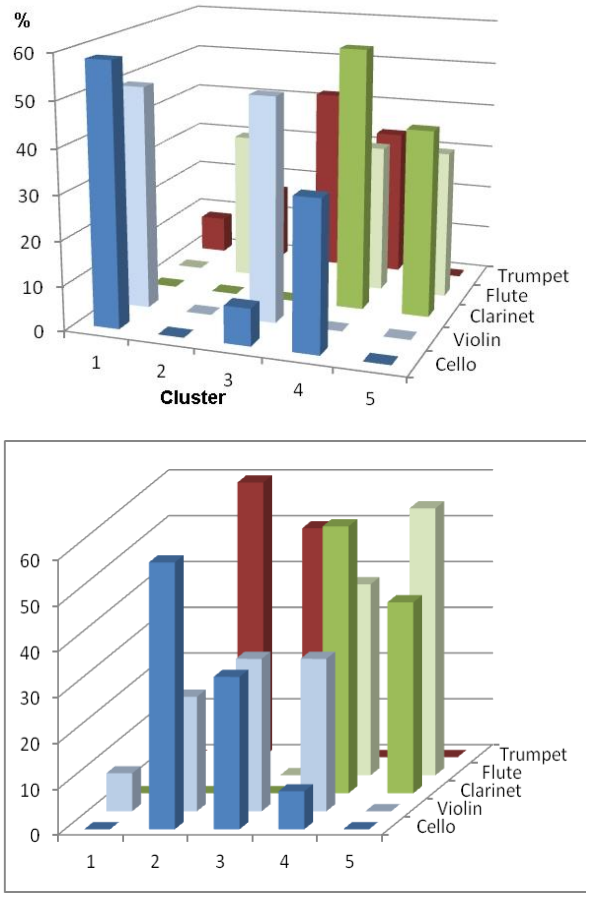

Figure 3. Clusterization by musical instruments in percentages of data by database Tinysol (Up) and Goodsounds (Bottom).

The ideal clustering values for K-means are usually given through the inflection points of the elbow method, represented in the graph of Inertia versus the number of

clusters (see Figure 4). Three inflections are observed, in K=2, K=4, and K=5; That would be the ideal cluster numbers to minimize variance if the clusters were all of equal size. The value K=2 would be used if a partition is desired to compare the two databases. The value K=5 corresponds to a partition in terms of the five musical instruments. The data used to suggest another partition in K=4 for families of Instruments; however, their acoustics invite us to consider only three different families: chordophones, wooden aerophones, and metal aerophones. Possibly the lack of data from other metal aerophones causes the inflection bias which would move from K=4 to K=3 if we had more metal aerophones common to the two databases used.

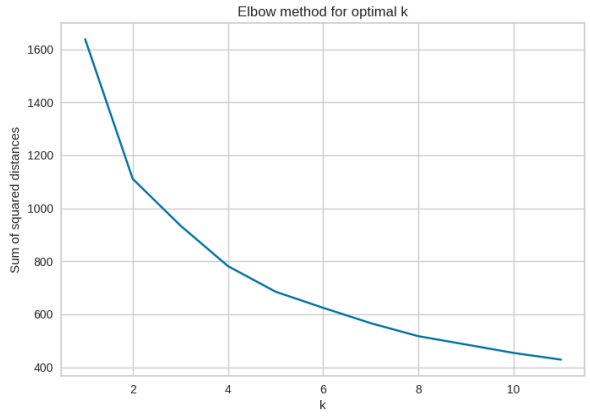

Figure 4. Elbow method for clustering audio records, both databases.

3.2 Clustering by Family of instruments

Figure 5 shows the clustering with K=3, which corresponds to an average separation between data and centroids of 10.82 (Inertia of 117). These values effectively discriminate between wooden and metal aerophones, since the former are significantly centered in cluster 2, and the trumpet only in clusters 2 and 3. The right part of Figure 5 shows that clustering is more effective for aerophones than chordophones, and in general, the data are effectively separated by more than 50%.

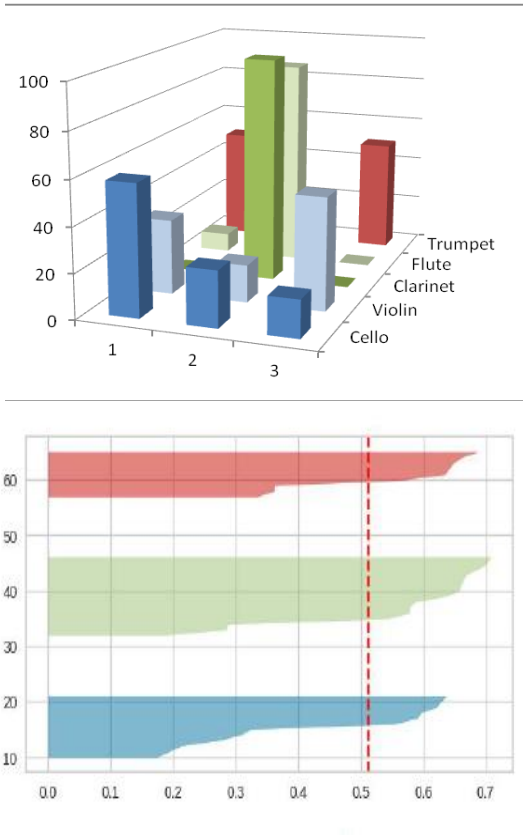

Figure 5. K-Means clustering by families of musical instruments in percentages of data per instrument (Up), Silhouette Diagram (Bottom).

3.3 Clustering by musical sounds

For each musical instrument, a K-means analysis was performed to evaluate the clusters by sound (K=12). Figures 6 and 7 show the results of the Clarinet (36 audios) and the Flute (60 audios). The color scale indicates the percentage of records in each cluster by musical sound. The results corroborate that given an instrument and musical sound, the characteristic timbral space occupies defined regions [5] and the tabulated values of its characteristic timbral coefficients [4] are within the cluster referred to in Figures 6, 7, and 8.

In each figures a clustering by musical notes is observed, since 12 sounds corresponding to the fourth musical octave of each instrument were considered, the total number of clusters is K=12, identified at the top of each figure with numbers ranging from 0-11.

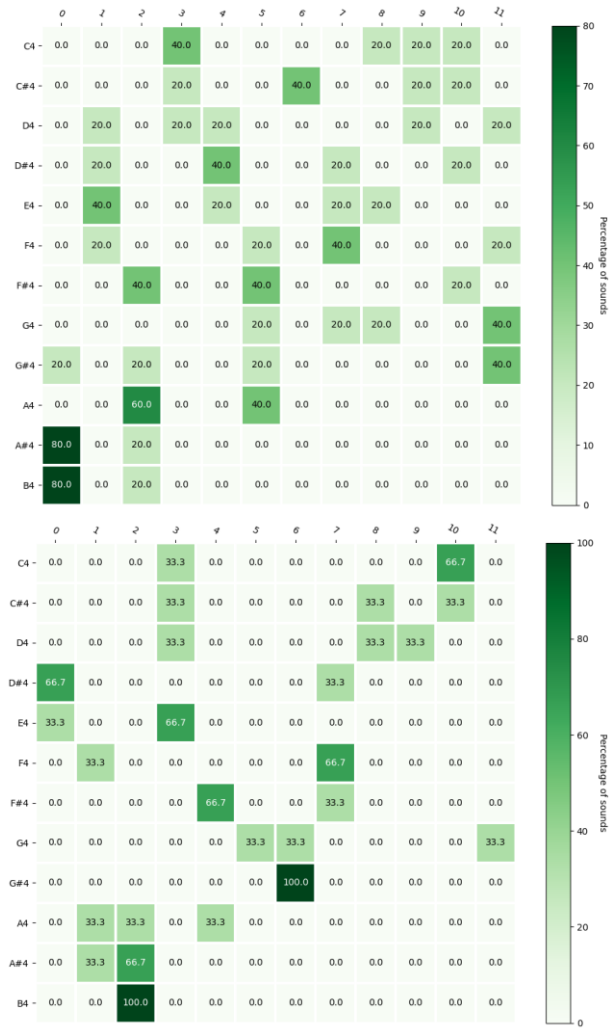

Figure 6. Matrix of clustering by musical notes for Flute (Up) and Clarinet (Bottom). The numbers at the top of the figures correspond to each of the 12 clusters (0 - 11).

It can be seen that each sound occupies a specific region, that is, it does not occupy the same set of clusters for another musical note. For example, it is observed that the total data of sound C4, for the clarinet, occupies two clusters 3 and 10 and there is no other musical note that occupies only these two clusters. This analysis is extended to all musical instruments and notes.

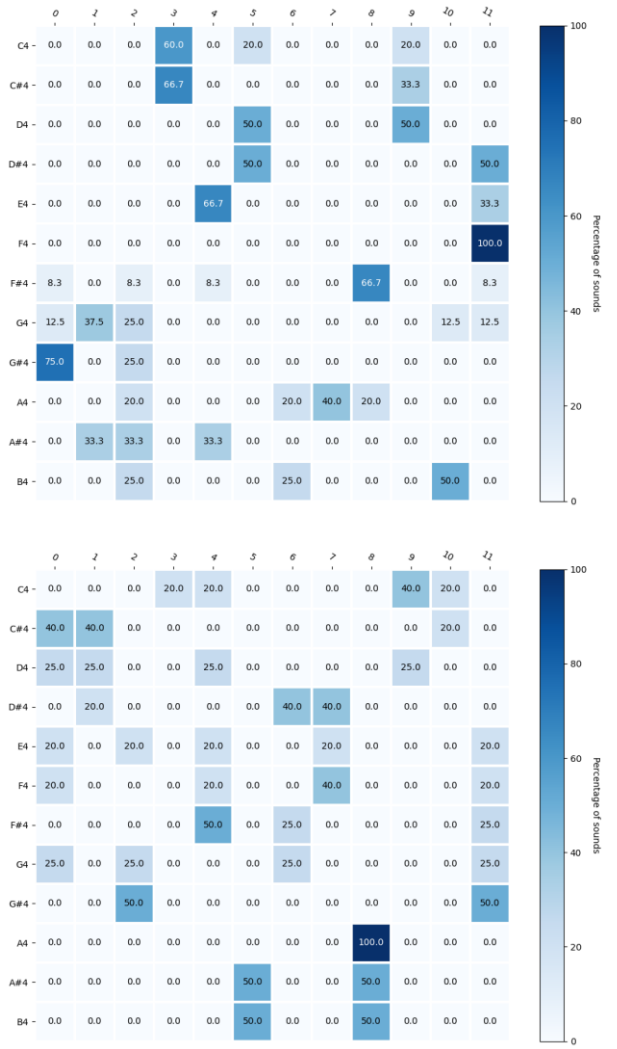

Figure 7. Matrix of clustering by musical notes for Violin (Up) and Cello (Bottom). The numbers at the top of the figures correspond to each of the 12 clusters (0 - 11).

Sounds of the same fundamental frequency (f_0) on different musical instruments can have timbral similarity, so their audio records should share the same region of timbral space. For each of the 12 fourth octave sounds, a K-means analysis was performed with K=5. As an example, Figure 9 shows the result for note D4, showing that the FFT have similarity in terms of the distribution of the partial components both in frequency and amplitude.

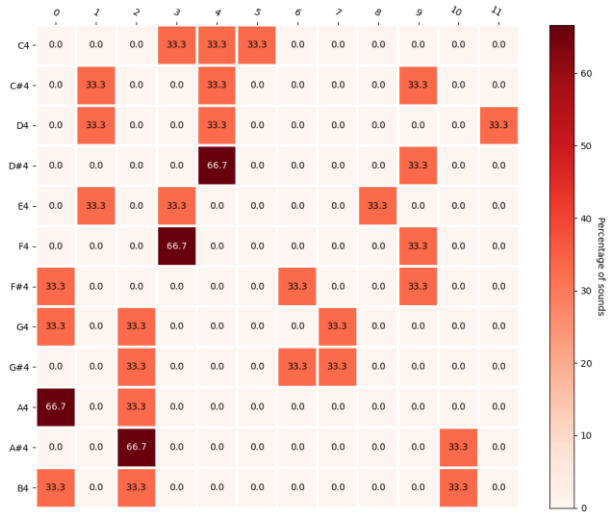

Figure 8. Matrix of clustering by musical notes for Trumpet. The numbers at the top of the figures correspond to each of the 12 clusters (0–11).

It is observed in the trumpet that no sound occupies a single cluster, being an element that differentiates this instrument from the group of musical instruments analyzed. In addition, it is observed that the number of clusters per musical sound is bounded between 2 and 3 clusters.

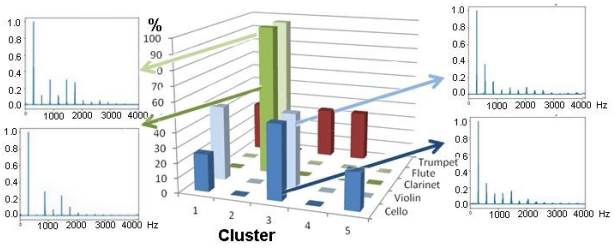

Figure 9. K-Mean clustering by families of instruments for the D4 sound. Note the similarity in the FFT spectra for clusters 2 and 3.

As another relevant example, the C#4 sound results are shown (Figure 10). We can see that clustering predicts timbral similarity and similarity in the FFT envelopes of acoustically different musical instruments, as in the case of the Violin and Clarinet.

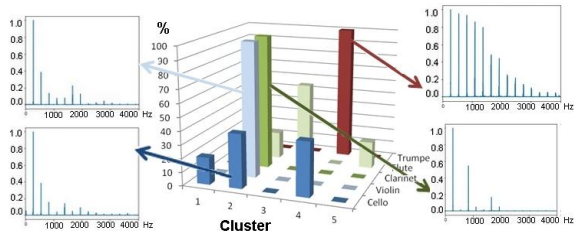

Figure 10. K-Mean clustering by instrument families for the C#4 sound. Note the similarity in the spectra of the FFTs in cluster 1 and the difference to the trumpet FFT in clusters 4.

4. CONCLUSIONS

The relevant aspects of musical timbre are contained in the Fast Fourier Transform, which can be characterized by a set of dimensionless coefficients (Affinity, Sharpness, Harmonicity, Monotony, Medium Contrast, and Medium Affinity) that together with the fundamental frequency form a timbral space of seven dimensions, where the timbral similarity can be defined as the geometric proximity through a Euclidean metric in that space.

The exploratory analysis of the Goodsounds and TinySol monophonic audio databases using the K-means algorithm allowed partial clustering of the data. The elbow criterion allowed us to define the groups of possible clusters for the analyzed data, K=2 to compare databases and K=5 for musical instruments (Figure 4).

In the partition by musical instruments, we can observe that the data of each musical instrument appear in very specific bounded regions of the timbral space. Thus, all the Flute and Clarinet sounds appear in two clusters (4 and 5), the Trumpet sounds only in clusters 2 and 3, and none of the Violin and Cello data appears in cluster 5 (Figure 2). The same data partition was performed considering each of the individual Tinysol and Goodsounds databases, observing a similar behavior for the Flute, Clarinet, Violin, and Cello. As for the Trumpet, it does not occupy cluster 5 in any of the databases. When considering the division by Families of instruments, we find clear discrimination between wooden and metal aerophones; and between aerophones in general and chordophones (Figure 5). Considering the total data, these are effectively separated by more than 50%. For the partition by musical sounds, the results corroborate that given an instrument and musical sound, the characteristic timbral space occupies defined regions. Thus, it was observed that no musical sound occupies the same cluster as another sound (Figure 6 - 8).

The inertia value for instrument partitioning states that the mean distance between the data and the center of the cluster is much less than the distances between each of the five clusters. A similar behavior occurs when we look at families of instruments. These results allow us to affirm that the distance between the groups of clusters analyzed is well-discriminated.

Finally, In the partition by musical notes for the timbral similarity analysis, it was possible to identify similarities in the distribution of the partial components (both in frequency and amplitude) between the FFT spectra when compared with another instrument, consistent with the hypothesis that timbrally similar sounds, in timbral space, must belong to the same cluster (Figure 9 - 10).

For future work, it is proposed to extend this analysis to the rest of the musical octaves, to incorporate other musical dynamics and other musical instruments.

5. ACKNOWLEDGMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

6. REFERENCES

[1] F. Simonetta, S. Ntalampiras and F. Avanzini, "Survey and future challenges," Multimodal music information processing and retrieval, In Intern. workshop on multilayer music representation Proc. (*MMRP*), pp.10 – 18, 2019.

[2] K. Tatar, D. Bisig and P. Pasquier, "Latent timbre synthesis: Audio-based variational auto-encoders for music composition and sound design applications", *Neural Computing and Applications*, Vol. 33, pp. 67-84, 2021.

[3] S. McAdams, "The perceptual representation of timbre", *Timbre: Acoustics, Perception, and Cognition*, Springer: Cham, Switzerland, pp. 23–57, 2019.

[4] M. Caetano, C. Saitis, K. Siedenburg, "Audio content descriptors of timbre", *Timbre: Acoustics, perception, and cognition*, Springer: Cham, Switzerland. pp. 297-333, 2019.

[5] G. Peeters, B.L. Giordano, P. Susini, et al "The timbre toolbox: Extracting audio descriptors from musical signals", *JASA*, vol. 5, pp. 2902-2916, 2011.

[6] Y. Gonzalez and R. C. Prati "Acoustic Descriptors for Characterization of Musical Timbre Using the Fast Fourier Transform", *Electronics*, vol. 11, no 9, pp. 1405, 2022.

[7] Y. Gonzalez and R. C. Prati. "Similarity of Musical Timbres Using FFT-Acoustic Descriptor Analysis and Machine Learning", *Eng,* vol. 4, no 1, pp. 555-568, 2023.

[8] Y. Gonzalez and R. C. Prati, "Acoustic Analysis of Musical Timbre of Wooden Aerophones", *RJAV* vol. 19, no 2, pp. 134-142, 2022.

[9] O. Romaní Picas, H. Parra-Rodriguez, D. Dabiri, et al."A real-time system for measuring sound goodness in instrumental sounds", *Proc. of the 138th Audio Engineering Soc. Conv*. (Warsaw, Poland) pp. 1106–1111, 2015.

[10] E. Carmine, D. Ghisi, V. Lostanlen et al. "*TinySOL: An Audio Dataset of Isolated Musical Notes*", Zenodo, 2020.

[11] P. Virtanen, R. Gommers, T.E. Oliphant, et al. "SciPy 1.0 Contributors SciPy 1.0 Fundamental Algorithms for Scientific Computing in Python", *Nat. Methods*, pp. 261 – 272, 2020.