Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 175 (2020) 96–101

Procedia Computer Science

www.elsevier.com/locate/procedia

The 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC)
August 9-12, 2020, Leuven, Belgium

Segmentation process and spectral characteristics in the determination of musical genres

Amelec Viloria [a]*, Omar Bonerge Pineda Lezama[b], Danelys Cabrera[c]

[a,b] Universidad de la Costa, Barranquilla, Colombia.
[c]Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras

Abstract

Over the past few years there has been a tendency to store audio tracks for later use on CD-DVDs, HDD-SSDs as well as on the internet, which makes it challenging to classify the information either online or offline. For this purpose, the audio tracks must be tagged. Tags are said to be texts based on the semantic information of the sound [1]. Thus, music analysis can be done in several ways [2] since music is identified by its genre, artist, instruments and structure, by a tagging system that can be manual or automatic. The manual tagging allows the visualization of the behavior of an audio track either in time domain or in frequency domain as in the spectrogram, making it possible to classify the songs without listening to them. However, this process is very time consuming and labor intensive, including health problems [3] which shows that "the volume, sound sensitivity, time and cost required for a manual labeling process is generally prohibitive. Three fundamental steps are required to carry out automatic labelling: pre-processing, feature extraction and classification [4]. The present study developed an algorithm for performing automatic classification of music genres using a segmentation process employing spectral characteristics such as centroid (SC), flatness (SF) and spread (SS), as well as a time spectral characteristic.

Keywords: Supervised learning algorithms; Music genres classification; Centroid (SC); Flatness (SF); Spread (SS).

* Corresponding author. Tel.: +57-3046238313.
E-mail address: aviloria7@cuc.edu.co

1877-0509 © 2020 The Authors. Published by Elsevier B.V.

10.1016/j.procs.2020.07.016

1. Introduction

In the present paper, a classification scheme was developed in which, initially the input signal is processed to reduce noise, then the signal is segmented, and its segments are processed by two characterization schemes, one in the frequency domain and the other one in the time domain [5].

The study included an audio segmentation [6] using basic characteristics such as zero cross rate (ZCR) in addition to the calculation of energy in a very short time "centroid", using 2.4s windows, where a 98% accuracy in the classification was reported. There are also developments in digital image processing [7] focusing on the spectrogram whose objective is multiclass classification, where the used classifier was the Support Vector Machine (SVM), obtaining results of 86% multiclass classification, where the system determines which class the audio signal under analysis belongs to. Finally, the SVM classifier, despite being a classifier created for binary classification, obtains very good results due to the previous extraction of characteristics in the audio track.

The study considers some characteristics proposed by Tzanetakis and Cook [8] were used, such as the spectral centroid (point where the spectrum is in equilibrium) and ZCR (average value of the times that the signal crosses zero in the x-axis in the time domain). In addition, other characteristics such as Spectral Flatness (value of the amount of frequency changes per frame) and Spectral Spread (power around each Spectral Centroid and the relationship of one centroid to the others) were used.

2. Proposed method

The proposed system is shown in Figure 1, which classifies a set of audio tracks divided into 5 genres. In particular, during the training, 10 audio tracks of each music genre were used, which were Cumbia, Pop, Rap, Rock and Salsa with sampling frequency fs = 44100Hz, 16 bits deep, with a duration of 5 - 10 minutes in wav format. Three spectral characteristics were extracted: Spectral Centroid, Spectral Spread and Spectral Flatness, as well as a temporal characteristic named Zero Crossing Rate. Finally, 4 classifiers were used: Decision Trees, Discriminant Analysis, Support Vector Machine (Gaussian), and Nearest Neighbor Classifier [9] [10] [11] [12] [13] [14].

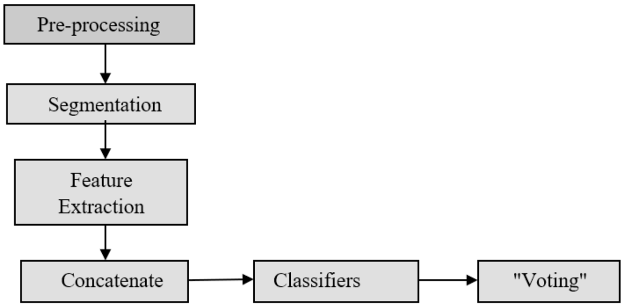

Fig. 1. Audio signal classification method.

2.1. Extraction of characteristics

As mentioned at the beginning of this chapter, four mathematical models will be used to obtain different characteristics of the audio track, one of them is in a time domain and three are spectral, so the FFT of the last three will have to be calculated.

- Zero Crossing Rate. This characterizer is of the temporal type, each instant of time is assigned a value obtained by a microphone called sample that has positive and negative values, which will be used to calculate the number of crossings by zero with equation (1) [15]:
-

$$ZCR = \frac{1}{2(N-1)} \sum_{m=1}^{N-1} |sgn[x(m+1)] - sgn[x(m)]|, \tag{1}$$

where x is the sample set, m is the sample position and N is the total of samples.

The objective of the algorithm is to add up the number of times a sample changes sign with respect to the previous one, meaning that the audio signal went from positive to negative values or vice versa in the x-axis, the values obtained are added up and the dividend is normalized by 2(N-1).

As it was observed in the segmentation section, 650 sub-segments were obtained with a length of 512 samples to which, applying the ZCR algorithm, 650 standardized components were obtained with 2 times the total of samples, that is, 1024 forming a characteristic vector with its respective label.

- Spectral Centroid. Before calculating the Fast Fourier Transform (FFT) and transforming the time domain values to the spectral domain, it is necessary to "window" the signal with an overlap, this to decrease the so-called "Gibbs" effect produced by abruptly cutting the signal. For this purpose, a Hamming window of 1024 (SW) size was used, which is twice the size of the number of samples to overlap by 50%. Thus, after "windowing" the signal and obtaining values with a size of 1024, FFT is applied to later calculate Spectral centroid with equation (2) [16]:

$$SC = \frac{\sum_{m=1}^{N-1} X(m) f(m)}{\sum_{m=1}^{N-1} X(m)}, \tag{2}$$

where X represents the values obtained from the FFT and f results from creating a vector of 1- 1024 values and dividing each value by 1024, a new scale is created representing f.

By calculating the centroids in the "sub-segment", another characteristic vector is obtained with a size of 650 values that resembles the vector obtained from ZCR; however, this characteristic represents the energy obtained in each window.

- Spectral Spread. It represents the concentration of energy around each Spectral centroid. An important feature is that the higher the SS, the larger the change in frequencies, calculated after having SC over the 1024 window using equation (3) [16]:

$$SS = \sqrt{\frac{\sum_{m=1}^{N-1} (f(m) - SC(m)) * |X(m)|}{\sum_{m=1}^{N-1} |X(m)|}}. \tag{3}$$

The only difference in the calculation of SS is that a subtraction is made to f with the SC that was obtained, in addition to calculating its square root in that window.

- Spectral Flatness. This feature belongs to the set of basic characteristics [16], which indicates how "flat" the spectrum is with a series of values expressing the energy of the spectrum within a predefined frequency band, equation (4) is concerned:

$$SF = \frac{\sqrt{\prod_{m=1}^{N-1} |X(m)|}}{\frac{1}{N} \sum_{m=1}^{N-1} |X(m)|}. \tag{4}$$

2.2. Concatenating vectors

4 characteristic vectors are obtained, whose size is 650. These vectors are concatenated and obtaining 1 vector of 2600 size; the order will be ZCR-SC-SS-SF-Label. This process will be completed with the 750 vectors, obtaining the final descriptor that will be classified by 4 methods [17][18].

2.3 Classifiers

The classification is made in each sub-segment whose duration will be 11.2 ms, that is to say that there will be 750 classifications and, as mentioned in the paper, 4 classifiers were used, obtaining the classification values shown in Table 1. In most cases the value obtained in the main diagonal of the confusion matrix is more than double and in the best case (rap) the trend is 6 times higher than the highest of the other 4 options. Furthermore, observing the values of the ROC curve of the cumbia genre (graphical representation of the comparison between sensitivity y-axis against specificity y-axis where the maximum value is 1) shows an excellent classification, see Figure 1. The value of the ROC curve for all genera is shown in Figure 2.

Table 1. Classification percentage of musical genres.

Classifiers	
Type	**Accuracy**
Decision Tree	34.04 %
Discriminant Analysis	51.6 %
Support Vector Machine	57.45 %
K-Nearest Neighbor	50.2 %

Table 2. Confusion matrix of musical genres.

Gaussian SVM					
	Cumbia	Pop	Rap	Rock	Salsa
Cumbia	72	42	7	9	25
Pop	16	82	25	4	27
Rap	14	17	110	2	19
Rock	13	4	7	90	35
Salsa	22	10	18	17	92

2.4 Voting

Once the classification of each sub-segment is obtained, that is, 15 classified vectors of the same song, the researcher proceed to choose the value that is most repeated within the 15 sub-segments as shown in Table 3.

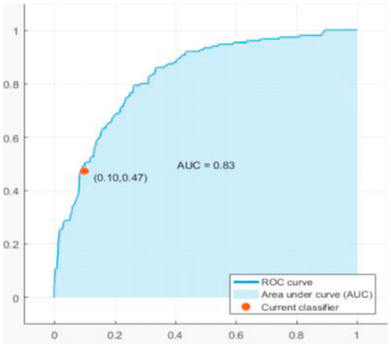

Fig. 2. ROC curve classification of musical genres.

Table 3. ROC curve classification of musical genres.

Genre	ROC curve value
Violin	0.984
Piano	0.971
Guitar	0.982
T-flute.	0.999

3. Tests and Results

A multiclass classification of 96% was obtained in the process, using the Gaussian SVM and the voting method with audio signals that have very similar spectral components. The same procedure was done for the classification of 10 songs played with 4 different musical instruments obtaining excellent results with the 5 classifiers where the SVM was also the classifier that provided the best results, as shown in Table 4. This can also be observed in the ROC curve of Figure 3.

Table 4. Classification percentage of musical instruments

Classifiers	
Type	**Accuracy**
Decision Tree	72.4 %
Discriminant Analysis	82.0 %
Support Vector Machine	94.3 %
K-Nearest Neighbor	77.5 %

Table 5. Confusion matrix for musical instruments.

Gaussian SVM				
	Violin	Piano	Guitar	T-flute.
Violin	148	2	3	4
Piano	0	140	11	4
Guitar	14	1	147	4
T-flute.	7	3	4	141

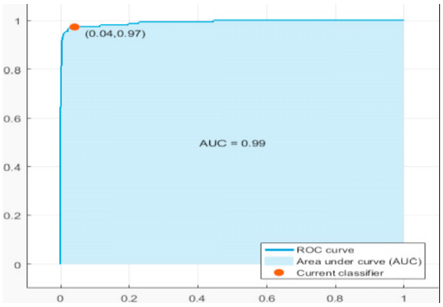

Fig. 3. ROC curve classification of musical instruments.

4. Conclusions

In order to make a correct classification, it is necessary to know the input vectors and to discriminate defects that they can have. In addition, it is observed that the classification depends to a great extent on how similar the vectors are (genres or musical instruments). In these cases, the two tests verified this idea although they were the same size, segmentation and extraction of characteristics, the values changed widely. The voting process is very efficient if there is a regular to good classification in several segments despite the low level obtained by SVM. When applying voting, the classification was 99% for musical instruments obtaining satisfactory results and observing that depending on the amount of frequencies involved (instruments and voice), the classification index will be lower, even when the voting method improves it notably.

References

[1] Viloria, A., Vargas, J., Cali, E. G., Sierra, D. M., Villalobos, A. P., Bilbao, O. R., … Hernández-Palma, H. (2020). Big Data Marketing During the Period 2012–2019: A Bibliometric Review. In Advances in Intelligent Systems and Computing (Vol. 1039, pp. 186–193). Springer. https://doi.org/10.1007/978-3-030-30465-2_21

[2] Mitrovic, D., Zeppelzauer, M., Eidenberger, H.: Analysis of the Data Quality of Audio Features of Environmental Sounds. Knowledge Creation Diffusion Utilization, pp. 4– 17 (2006)

[3] Juthi, J. H., Gomes, A., Bhuiyan, T., & Mahmud, I. (2020). Music Emotion Recognition with the Extraction of Audio Features Using Machine Learning Approaches. In Proceedings of ICETIT 2019 (pp. 318-329). Springer, Cham.

[4] Greece-Duan, S., Zhang, J., Roe, P.: A survey of tagging techniques for music, speech and environmental sound, pp. 637–661 (2014)

[5] Lee, C. S., Tsai, Y. L., Wang, M. H., Sekino, H., Huang, T. X., Hsieh, W. F., ... & Yamaguchi, T. (2019, November). FML-based Machine Learning Tool for Human Emotional Agent with BCI on Music Application. In 2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI) (pp. 1-6). IEEE.

[6] Rana, D., & Sandhu, R. (2019). Music Recommendation System using Machine Learning Algorithms.

[7] Faisal-Ahmed, P.P., Paul, M.G.: Music Genre Classification Using a Gradiente-Based Local Texture descriptor. Springer International Publishing Switzerland, pp. 99–110 (2016)

[8] Tzanetakis, G.: Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, pp. 293–302 (2002)

[9] Munkhbat, K., & Ryu, K. H. (2020). Classifying Songs to Relieve Stress Using Machine Learning Algorithms. In Advances in Intelligent Information Hiding and Multimedia Signal Processing (pp. 411-417). Springer, Singapore.

[10] Duarte, A. E. L. (2020). Algorithmic interactive music generation in videogames. SoundEffects-An Interdisciplinary Journal of Sound and Sound Experience, 9(1), 38-59.

[11] Finley, M., & Razi, A. (2019, January). Musical Key Estimation with Unsupervised Pattern Recognition. In 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0401-0408). IEEE.

[12] Pelchat, N., & Gelowitz, C. M. (2019, May). Neural Network Music Genre Classification. In 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE) (pp. 1-4). IEEE.

[13] Choi, J., Lee, J., Park, J., & Nam, J. (2019). Zero-shot learning for audio-based music classification and tagging. arXiv preprint arXiv:1907.02670.

[14] Ahuja, M., & Sangal, A. L. (2018, December). Opinion Mining and Classification of Music Lyrics Using Supervised Learning Algorithms. In 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC) (pp. 223-227). IEEE.

[15] Calvo-Zaragoza, J., Micó, L., & Oncina, J. (2016). Music staff removal with supervised pixel classification. International Journal on Document Analysis and Recognition (IJDAR), 19(3), 211-219.

[16] Schreiber, H., & Müller, M. (2017). A Post-Processing Procedure for Improving Music Tempo Estimates Using Supervised Learning. In ISMIR (pp. 235-242)..

[17] Benavides, E. S., Charris, F. C., & Viloria, A. (2020). Inequality in Writing Competence at Higher Education in Colombia: With Linear Hierarchical Models. In Advances in Intelligent Systems and Computing (Vol. 1039, pp. 122–132). Springer. https://doi.org/10.1007/978-3-030-30465-2_15

[18] Viloria, A., Lis-Gutiérrez, J. P., Gaitán-Angulo, M., Godoy, A. R. M., Moreno, G. C., & Kamatkar, S. J. (2018). Methodology for the design of a student pattern recognition tool to facilitate the teaching - Learning process through knowledge data discovery (big data). In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 10943 LNCS, pp. 670–679). Springer Verlag. https://doi.org/10.1007/978-3-319-93803-5_63