VOL. 20, NO. 1, 2021, 43-53
www.elektrika.utm.my
ISSN 0128-4428

ELEKTRIKA
Journal of Electrical Engineering

Statistical Analysis of the Sonic Effects of Music from a Comprehensive Datasets on Audio Features

Tobechukwu Okechukwu Otuokere[1], **Agbotiname Lucky Imoize**[2,3*] and **Aderemi A. Atayero**[4]

[1]Department of Computer Science, National Open University of Nigeria, Victoria Island, Lagos, Nigeria
[2]Department of Electrical and Electronics Engineering, University of Lagos, Akoka 100213, Lagos, Nigeria
[3]Department of Electrical Engineering and Information Technology, Institute of Digital Communication, Ruhr University, 44801 Bochum, Germany.
[4]Department of Electrical and Information Engineering, Covenant University, Ota 112233, Ogun State, Nigeria

[*]Corresponding author: aimoize@unilag.edu.ng

Abstract: Music, for the longest time, has impacted human lives tremendously. The ability of music to access and activate a wide range of human emotions is sensational. Audio features provide various information necessary for sound engineers, music producers, and artists to improve their craft to excite the vast majority of music listeners globally. In this paper, an analysis of audio features derived using the Spotify web API endpoint and Spotify (Python module for Spotify web servers) is presented. The dataset was curated from audio features of over 160,000 songs released from year 1921-2020. For clarity, statistical descriptions and probability distribution functions of the audio features are presented. Additionally, the interrelationship and correlation amongst the various audio features are demonstrated. Overall, the dataset would find useful applications in classical and contemporary music production.

Keywords: Audio features, Music, Acousticness, Speechiness, Instrumentalness, danceability

Article History: received 12 September 2020; accepted 22 April 2021; published 30 April 2021.

1. INTRODUCTION

Audio features data provide essential information about how different subtle aspects of music relate to each other [1], [2]. It is not surprising that some of these features are not so often talked about in the literature. As audio technology continues to advance in the coming years, it is imperative to note the changing variables in the music industry, primarily as it affects the consumers [3], [4]. This, in turn, has enormous potentials to induce strategic business decisions in the music production process. Toward this end, this study presents audio features using the Spotify Web API endpoint and Spotify (Python module for Spotify web servers). The dataset reported was curated from the audio features of over 160,000 songs released from the year 1921-2020. The parameters investigated comprise audio features such as duration, key, energy, danceability, liveness, notation, mode, time signature, acousticness, instrumentalness, loudness, speechiness, valence, tempo, id, type, popularity.

This study analyzes the sonic effects of music from a comprehensive dataset on audio features for an informed decision by the various stakeholders in the music industry. These audio features are first curated and rigorously analyzed. The analysis was carried out in Jupyter notebooks with Python 3, using third-party packages-Pandas and Numpy as analytical tools. In the study, several audio feature parameters were investigated and analytically presented in section 4 of this paper.

The analyzed audio features will be of immense benefit to 1) Music producers and sound engineers in terms of sonic features to tweak for optimal results. 2) Artists and songwriters concerning the structure and "wordiness" of their songs [5]. The data will also provide further insights into how the most successful artists have inspired their audience [6]. Finally, the insights provided by this data will provoke further research and development in the field of music production and distribution.

The remaining part of this paper is organized as follows. Section 2 reports the related work and theoretical background. Section 3 presents the methodology, while Section 4 presents the results and useful discussions. Finally, the conclusion to the paper and future perspectives are given in Section 5.

2. RELATED WORK

In the existing works of literature, several studies have been reported on audio features analyses [7], [8], [9], [10], [11], [12], [13], [14]. In particular, the MARSYAS framework for audio analysis is proposed in [7]. The work reported a new method for temporal segmentation leveraging audio texture. In [8], audio content analysis was used to develop an approach to automatic segmentation and classification of audiovisual data. An audio stream can be classified and segmented into speech, music, environment sound, and silence using a suitable approach proposed in [9]. Additionally, the work introduced a set of new features comprising of the noise frame ratio and band periodicity.

The work in [10] proposes a multi-purpose approach capable of performing unsupervised audio analysis and

using an audio classification framework for audio features analysis. In [11], audio features were analyzed and evaluated for multitrack music mixtures. Additionally, the work introduces a convolutional neural network operating on a large temporal input for robust audio feature processing. Similarly, the lyrical and audio features of a song have been used to detect the emotion of the song [12]. An open-source python library called pyAudioanalysis was proposed in [13]. The library keeps track of several procedures for audio analysis. These comprise feature extraction, segmentation, audio signals classification, and content visualization. In [14], AENet, a new deep network was developed for audio event recognition.

Additionally, musical genre classification of audio signals was presented in [15], and a comprehensive definition of audio features for music content description is reported in [16]. A novel audio feature for music emotion recognition is presented in [17]. Also, audio features for noisy sound segmentation are elaborated in [18], and music genre classification using MIDI and audio features are outlined in [19]. In [11], extensive analysis and evaluation of audio features for multitrack music mixtures are presented.

In [20], features for content-based audio retrieval were examined, and the work in [21] takes a close look at semantic annotation and retrieval of music and sound effects. Furthermore, music and audio content are elucidated in [22], and the authors in [23] capture the modeling timbre distance with temporal statistics from polyphonic music. On the Integration of text and audio features for genre classification in music information retrieval [24], the authors present a clear exposition on how text and audio features can be integrated for optimal genre classification in the context of information retrieval.

On exploring automatic music annotation with "acoustically-objective" tags [25], over 10,000 songs curated from the Swat10k data set were annotated with a relatively large mixed vocabulary of over 600 tags. The authors compare short-time audio features and demonstrated that ENT features perform better than MFCC features. They are of the same dimensionality based on the GMM classifier. In another related study on the representations of sound in deep learning of audio features from music, the authors [26] applied a deep convolutional neural network (DCNN) to a relatively sizeable audio dataset. They adopted empirical methods performance the proposed algorithm on audio classification tasks. The authors' trained network shows robust performance in terms of the classification tasks when roughly 5s of music derived by less complex transformations of the raw audio waveform is fed as input. Furthermore, the authors observed that the highly structured spectrograms results from the STFT lacked precision when used for classification compared to the representation achieved by the random matrix transform of raw waveforms.

The subject of timbre-invariant audio features for harmony-based music is discussed in [27]. The authors present a new method geared towards enhancing the Chroma features. This is achieved by increasing the degree of timbre invariance while maintaining the features' discriminative power at the optimal threshold. The study also revealed that the authors trashed the lower coefficients while the upper coefficients were preserved. This is based on the idea that the lower Mel-frequency cepstral coefficients (MFCCs) are closely related to timbre.

Finally, score-independent audio features were adopted to describe music expression clearly in [28]. When music is being performed, the musician tends to add expressiveness to the musical message by varying the timing, dynamics, and timbre of the musical events to communicate an expressive intention in the most understandable manner. In the traditional context, music expression is analyzed following the acoustic parameters' deviations concerning the written score. The authors also demonstrated how machine learning could better understand expressive communication and derive audio features at an intermediate level.

In the next section, the methodology used for the audio features analysis is presented.

3. METHODOLOGY

In this paper, an analysis of a robust dataset comprising audio features extracted from over 160,000 songs released from 1921-2020 is presented. This was done in Jupyter notebooks with Python 3, using third-party packages; Pandas, and Numpy as analytical tools. The datasets reported in this article are outlined as follows, and a brief description of all the columns in the dataset is given in Table 1.

"data.csv" – Full dataset
"data_by_genres.csv" – Data by genres
"data_by_artist.csv" – Data by artists
"data_by_year.csv"- Data by year

The data source location is the Spotify Web API, and the Spotify Web API is based on REST principles. Data resources are accessed via standard HTTPS requests in UTF-8 format to an API endpoint, and Web API uses the appropriate HTTP verbs for each action.

First, we imported the essential libraries and dataset through several code lines and provided brief descriptive statistics of the dataset following some code lines in Python. Some columns that were not very useful were dropped using a few code lines and Python to ease data preprocessing. For suitable scaling, it was essential to convert the 'duration_ms' column to minutes. The distribution of sonic features in the dataset was critically examined, and the correlation between the audio features in the dataset was demonstrated to establish key interrelationships amongst the variables.

In terms of popularity, specific sonic features were considered, and a correlation plot was made to show the relationships between the said audio features and popularity. Also, the trends in audio features based on a time series were established. To this effect, a function to find the average of audio features per year was created, and a line chart showcasing how audio features have changed over time was presented. Furthermore, analysis of the oldies' audio features (songs before the year 2000) and Millenials (songs after the year 2000 till date) was reported.

A function to create a horizontal bar plot showcasing artists by popularity was first written, followed by showcasing the most famous artists before 2000 and distributing the sonic features. Finally, the most famous artists after the year 2000 and the sonic features'

distribution were reported. The dataset is analyzed and made available for research reproduction as comprehensive results reported in section 4 of this paper.

Table 1. Brief description of the audio features tested in the dataset

Audio Feature	**Description**
duration_ms	This audio feature describes the duration of the track in milliseconds.
key	This refers to the estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D ♭ , 2 = D, and so on. If no key was detected, the value is -1.
mode	The mode indicates the modality (major or minor) of a track and the type of scale from which its melodious content is derived. This is mainly represented by one, and the minor is 0.
time_signature	This describes an estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence; the track is acoustic. The distribution of values for this feature is referred to as the acousticness of the distribution.
danceability	Danceability describes how suitable a track is for dancing based on a combination of musical elements, including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is the least danceable, and 1.0 is the most danceable. The distribution of values for this feature is called the danceability distribution.
energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy. The distribution of values for this feature is often referred to as the energy distribution.
instrumentalness	The instrumentalness predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0. The distribution of values for this feature describes the instrumentalness distribution.
liveness	This detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides a strong likelihood that the track is live. The distribution of values for this feature is referred to as liveness distribution.
loudness	This audio feature explains the overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing the relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 dB. The distribution of values for this feature is called loudness distribution.
speechiness	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g., talk show, audiobook, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 represent tracks that may contain both music and speech.
valence	A measure from 0.0 to 1.0 is describing the musical positivity conveyed by a track. Tracks with high valence sound more positive (e.g., happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g., sad, depressed, angry). The distribution of values for this feature is referred to as the valence distribution.
tempo	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, the tempo is the speed or pace of a given piece and derives directly from the average beat duration. The distribution of values for this feature is called tempo distribution.
id	The Spotify ID for the specific track. This ID is assigned to each track to ease identification.
type	The object type: “audio_features.”
popularity	This audio feature gives useful information about the popularity of the track. The value will be between 0 and 100, with 100 being the most popular. The popularity is calculated by a suitable algorithm and is based, for the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have higher popularity than songs played a lot in the past. Artist and album popularity is derived mathematically from track popularity. Note that the popularity value may lag actual popularity by a few days: the value is not updated in real-time.

4. RESULTS AND DISCUSSIONS

In this section, the results and discussions are reported. Specifically, the results are presented in section 4.1 and the discussion of results is given in section 4.2.

4.1. Results of the Study

The results from the analyses are briefly described as follows. First, an overview of the dataset is shown in Table 2. Descriptive statistics about the dataset are given in Table 3. The distributions between audio features are reported in Fig. 1. The correlation between audio features in the dataset is presented in Fig. 2 and Fig. 3, respectively. In terms of popularity, Fig. 4 shows the correlation with some selected audio features, whereas Table 4 showcases the average of some selected audio features by year. Fig. 5 illustrates how some of these audio features have changed over the years, and Fig. 6 shows top artists of all time based on popularity. Fig. 7 shows the most famous artists before 2000 (Oldies), while Fig. 8 showcases audio features' distribution before 2000 (Millennials). Furthermore, Fig. 9 shows the most popular artists after the year 2000, and Fig. 10 shows the distribution of audio features after the year 2000. Finally, Fig. 11 gives the distribution of audio features after the year 2015 (Millennials).

4.2. Discussion of Results

The results of this study are briefly discussed as follows. In Table 2, Sergei Rachmaninoff, James Levine, and Berliner Philharmoniker recorded a duration of 831667, the energy of 0.211, and danceability of 0.279 each, while John McCormack had a duration of 159507, energy of 0.203, and danceability of 0.518. These results show that the John McCormack duration is lower, but the energy has a higher value. Following the results for a count across key audio features of 168592, in Table 3, the acousticness, danceability, and energy recorded mean values of 0.501360147, 0.533648407, and 0.48857702, respectively. Similarly, the standard deviations of the acousticness, danceability, and energy are 0.377992926, 0.175918949, 0.267346249, respectively. The maximum values of acousticness, danceability, and energy are 0.996, 0.988, and 1, respectively. These results imply that these features are highly important and each achieves nearly 100% measure, with the energy as most efficient, followed by the acousticness.

On the distribution of the audio features, as shown in Figure 1, the acousticness assumes a U-shaped distribution, while the danceability showed a reasonably normal distribution with a peak value of 0.60. For this distribution, the duration lies less than 12 minutes. The energy also gave a highly promising distribution, whereas the loudness did not pick until at -40 to 0. In this scenario, popularity is roughly 3000 but suddenly falls to around 300 till the 20th year. It rises slowly achieving a peak of 600 at the 45th year, and then fell steadily to 0 in the 90th year. The tempo is evenly distributed around 100 with a peak value of over 1000. Last, the valence was seen to almost occupy the entire range 0 to 1, and the years 1950-2020 showed a flat distribution with a peak at 4000.

Table 2. Summarized overview of the dataset

UNNAMED: 0	ACOUSTICNESS	ARTISTS	DANCEABILITY	DURATION_MS	ENERGY	EXPLICIT
0	0.732	['Dennis Day']	0.819	180533	0.341	0
1	0.982	['Sergei Rachmaninoff', 'James Levine', 'Berliner Philharmoniker']	0.279	831667	0.211	0
2	0.996	['John McCormack']	0.518	159507	0.203	0
3	0.982	['Sergei Rachmaninoff', 'James Levine', 'Berliner Philharmoniker']	0.279	831667	0.211	0
4	0.957	['Phil Regan']	0.418	166693	0.193	0

Table 3. Summarized view of the descriptive statistics of key audio features in the dataset

	UNNAMED: 0	ACOUSTICNESS	DANCEABILITY	DURATION_MS	ENERGY
COUNT	168592	168592	168592	168592	168592
MEAN	84295.5	0.501360147	0.533648407	232701.5574	0.48857702
STD	48668.46263	0.377992926	0.175918949	122392.1252	0.267346249
MIN	0	0	0	5108	0
25%	42147.75	0.0978	0.412	172160	0.265
50%	84295.5	0.515	0.543	209133	0.48
75%	126443.25	0.896	0.662	263707	0.709
MAX	168591	0.996	0.988	5403500	1

VOL. 20, NO. 1, 2021, 43-53
www.elektrika.utm.my
ISSN 0128-4428

ELEKTRIKA
Journal of Electrical Engineering

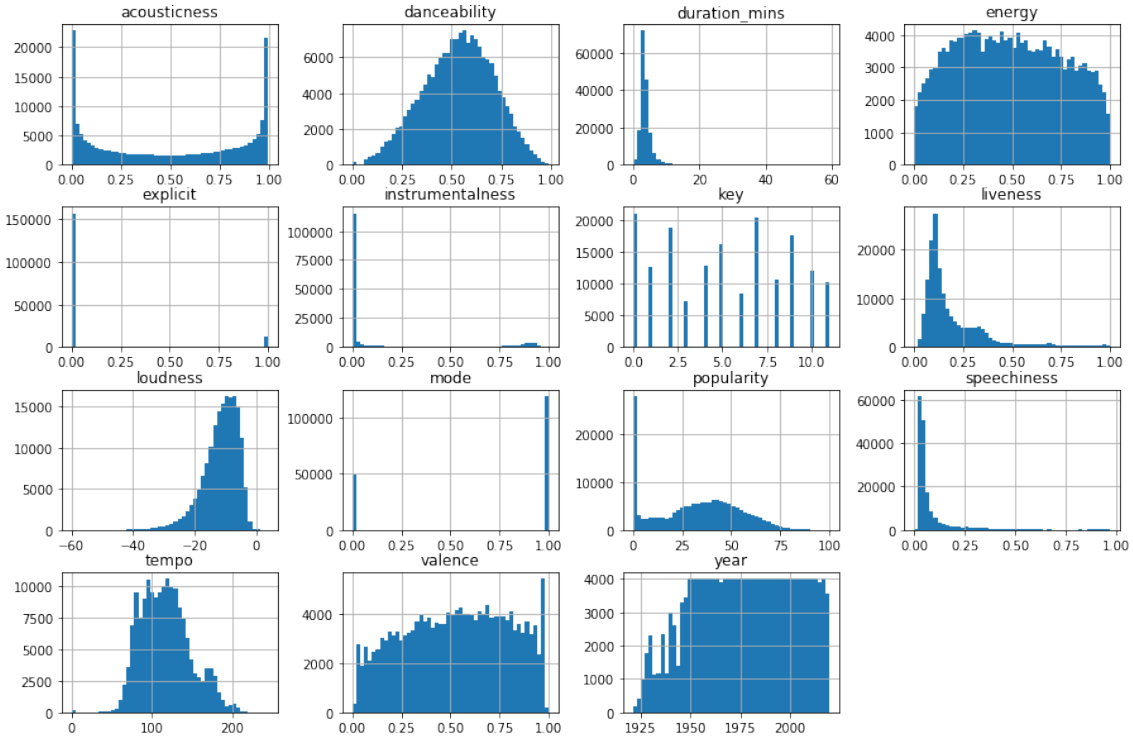

Figure 1. Distribution of the audio features

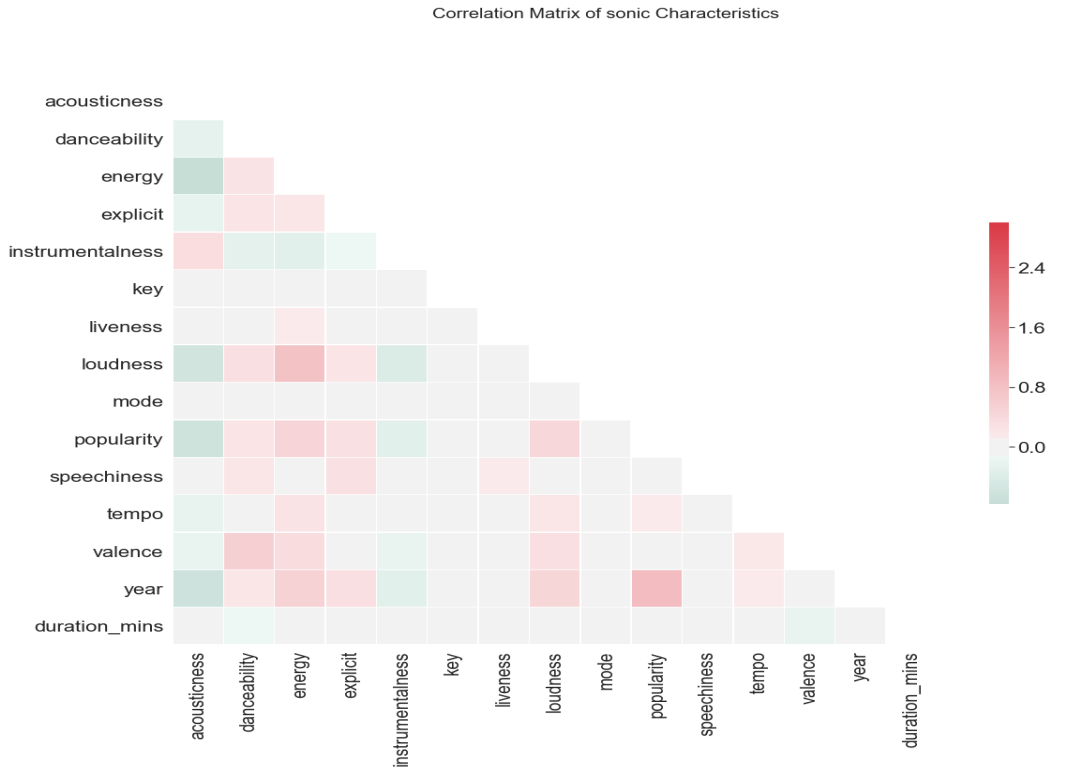

Fig. 2. Correlation matrix of sonic characteristics among the audio features

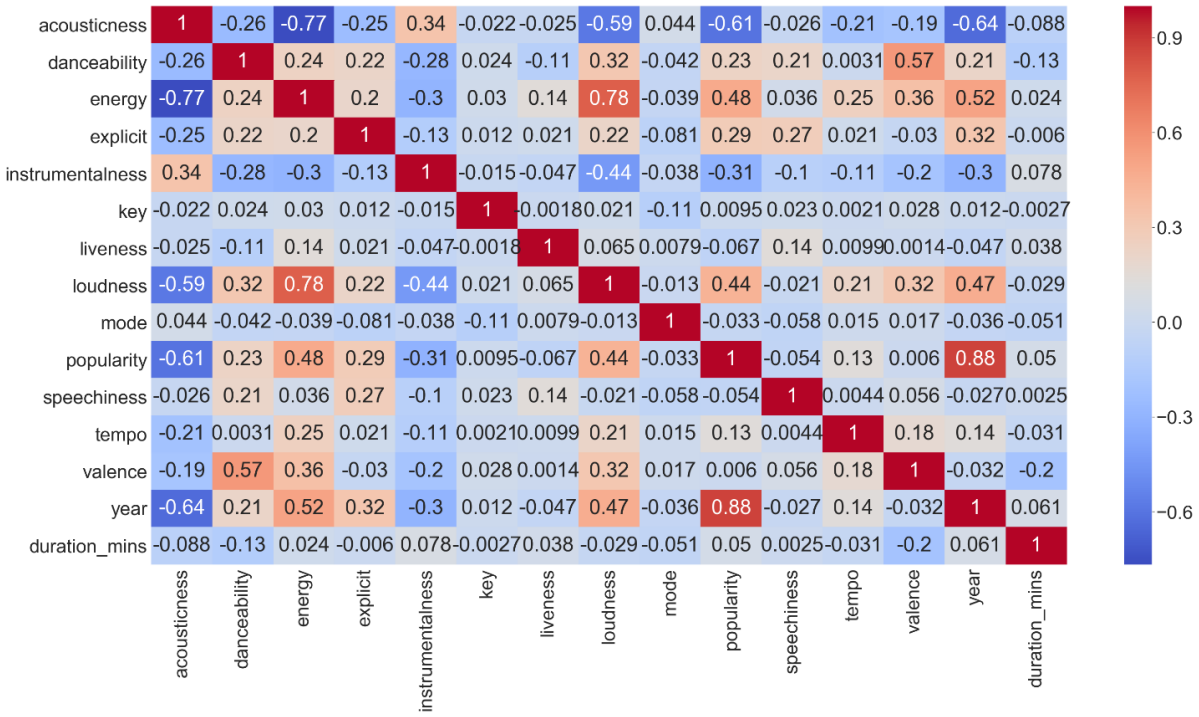

Figure 3. Detailed correlation matrix among the audio features investigated

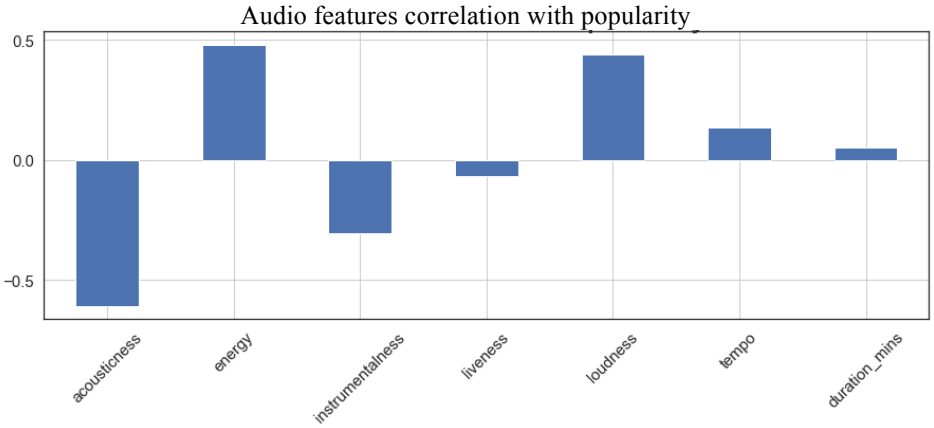

Fig. 4. Correlation between some selected audio features and popularity

Table 4. Average of some selected audio features by year

YEAR	ACOUSTICNESS	ENERGY	INSTRUMENTALNESS	LOUDNESS	DURATION_MINS	DANCEABILITY	EXPLICIT
1921	0.895823	0.236784	0.32233	-17.0954	3.831865	0.425661	0.054688
1922	0.939236	0.237026	0.44047	-19.18	2.798409	0.48	0
1923	0.976329	0.246936	0.401932	-14.3739	2.972605	0.568462	0
1924	0.935369	0.348118	0.58281	-14.1567	3.081525	0.549403	0
1925	0.965422	0.264373	0.408893	-14.5167	3.068845	0.57189	0

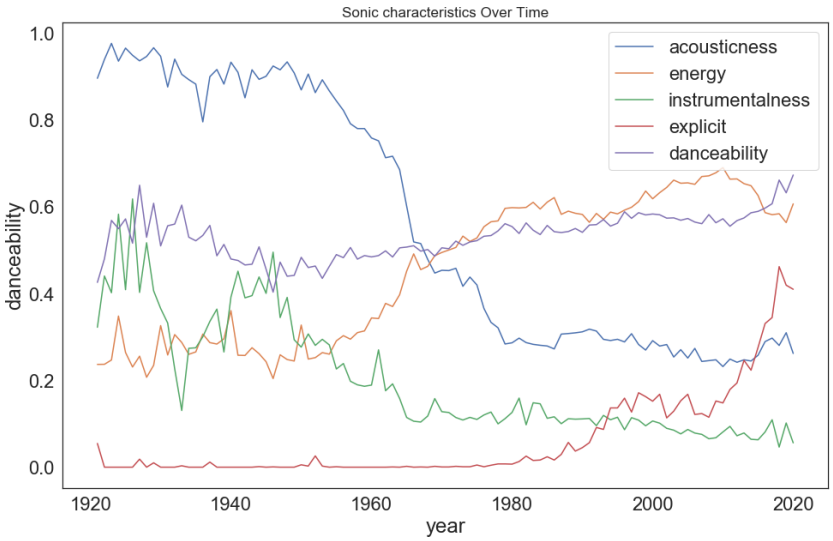

Figure 5. Changes in audio features over time in years

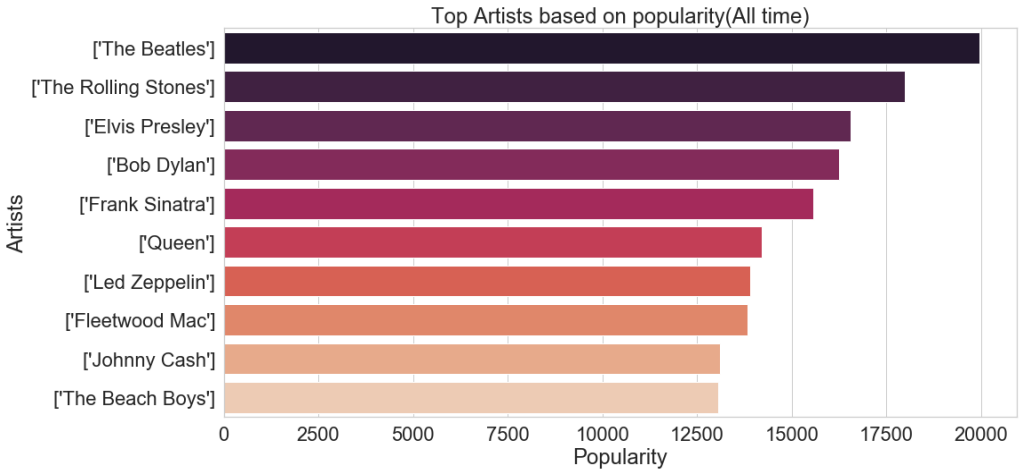

Figure 6. Top artists (all time) based on popularity

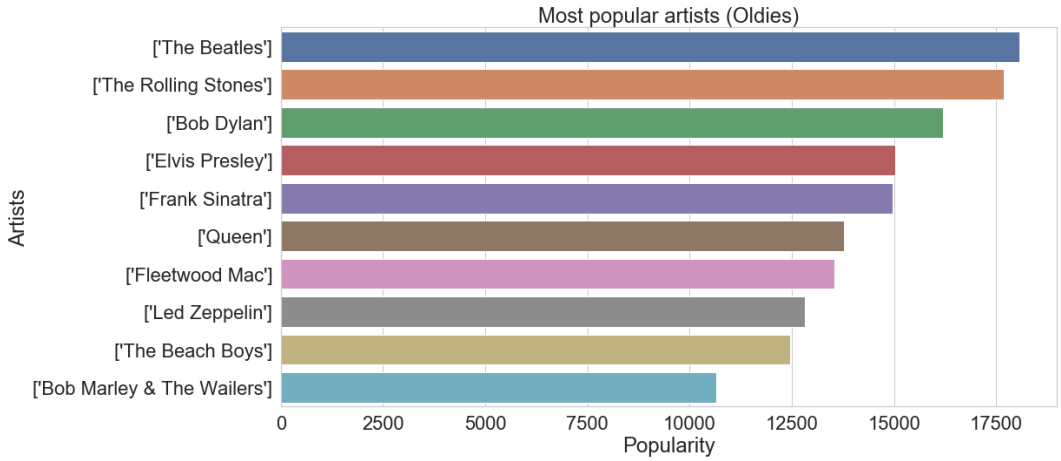

Figure 7. Most popular artists (oldies)

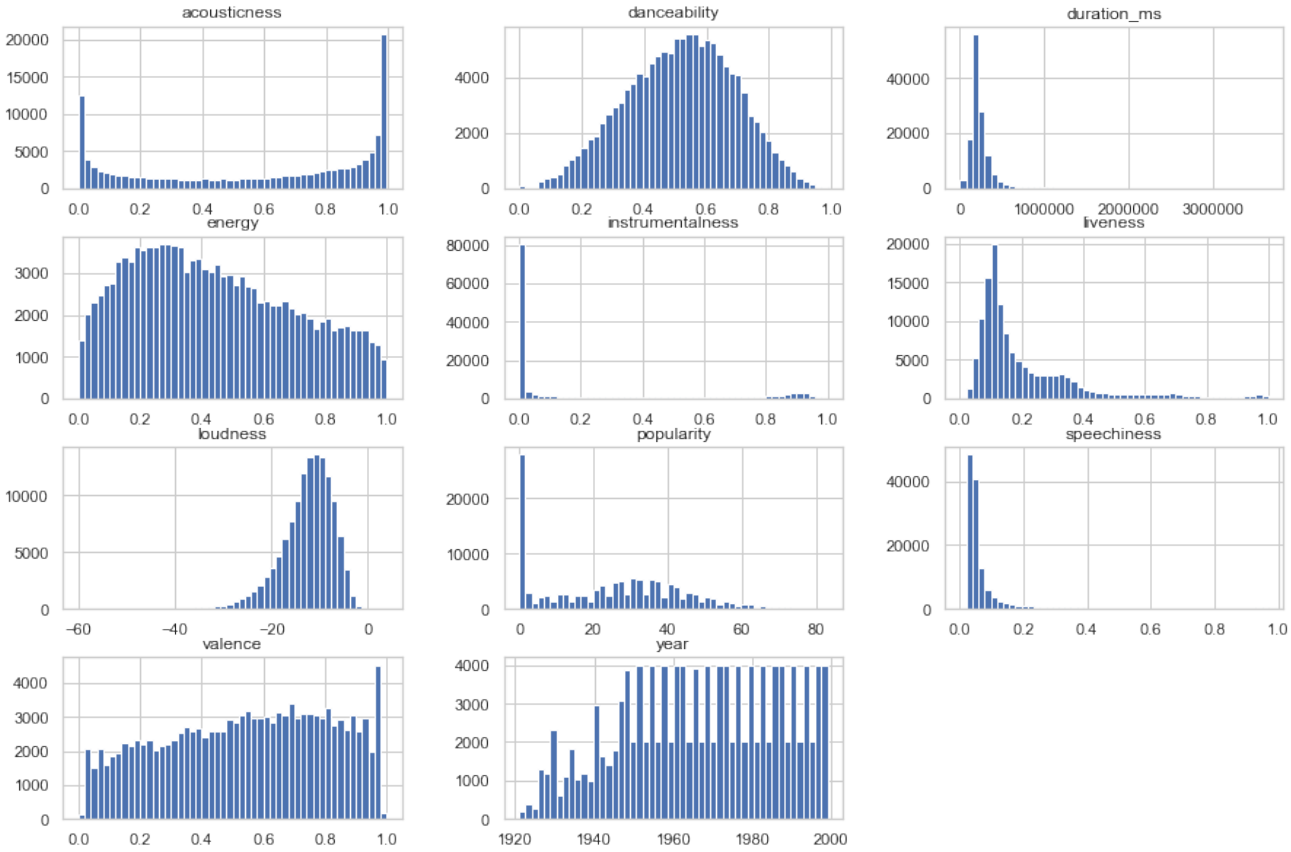

Figure 8. Distribution of audio features before the Year 2000

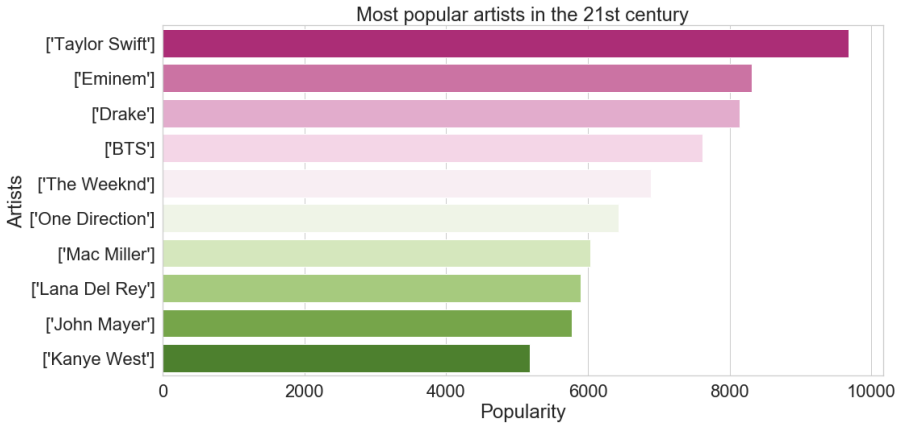

Figure 9. Most popular artists after the year 2000 (Millennials)

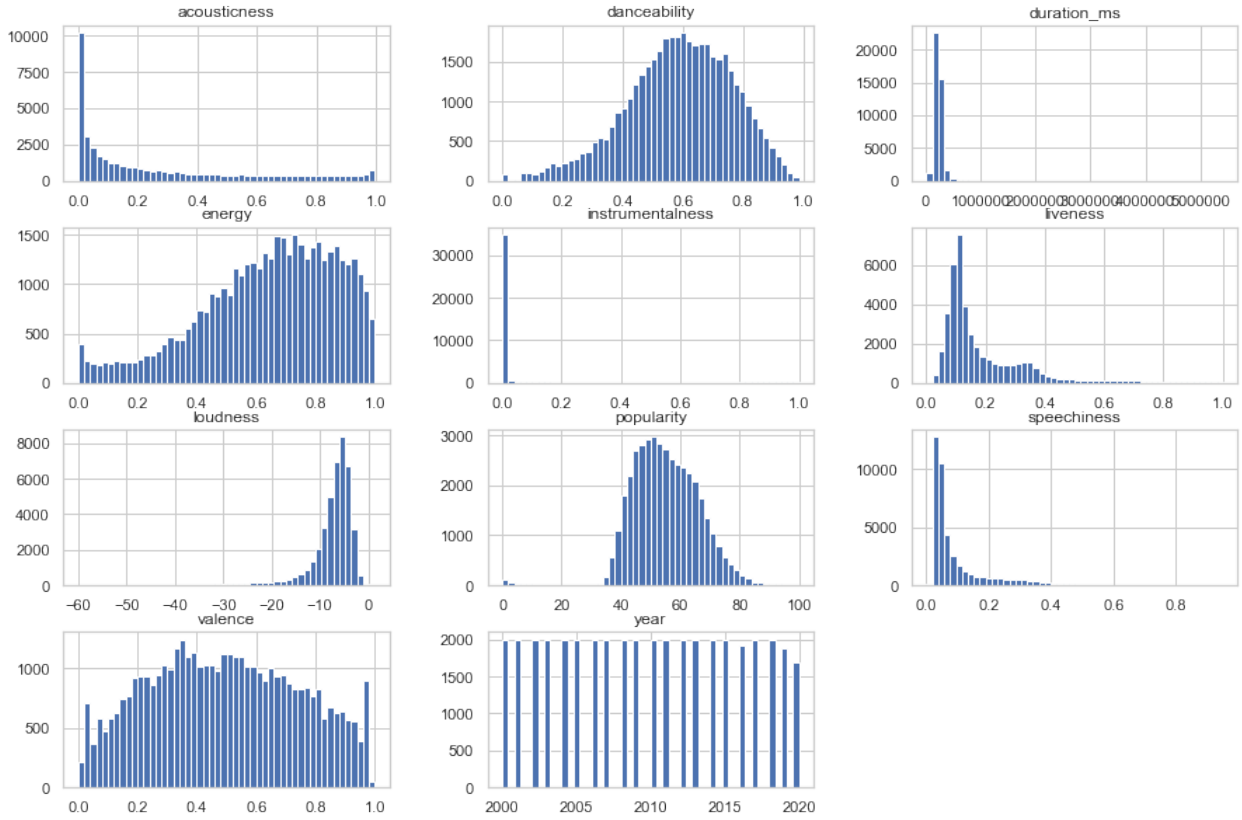

Figure 10. Distribution of audio features after the year 2000 (Millennials)

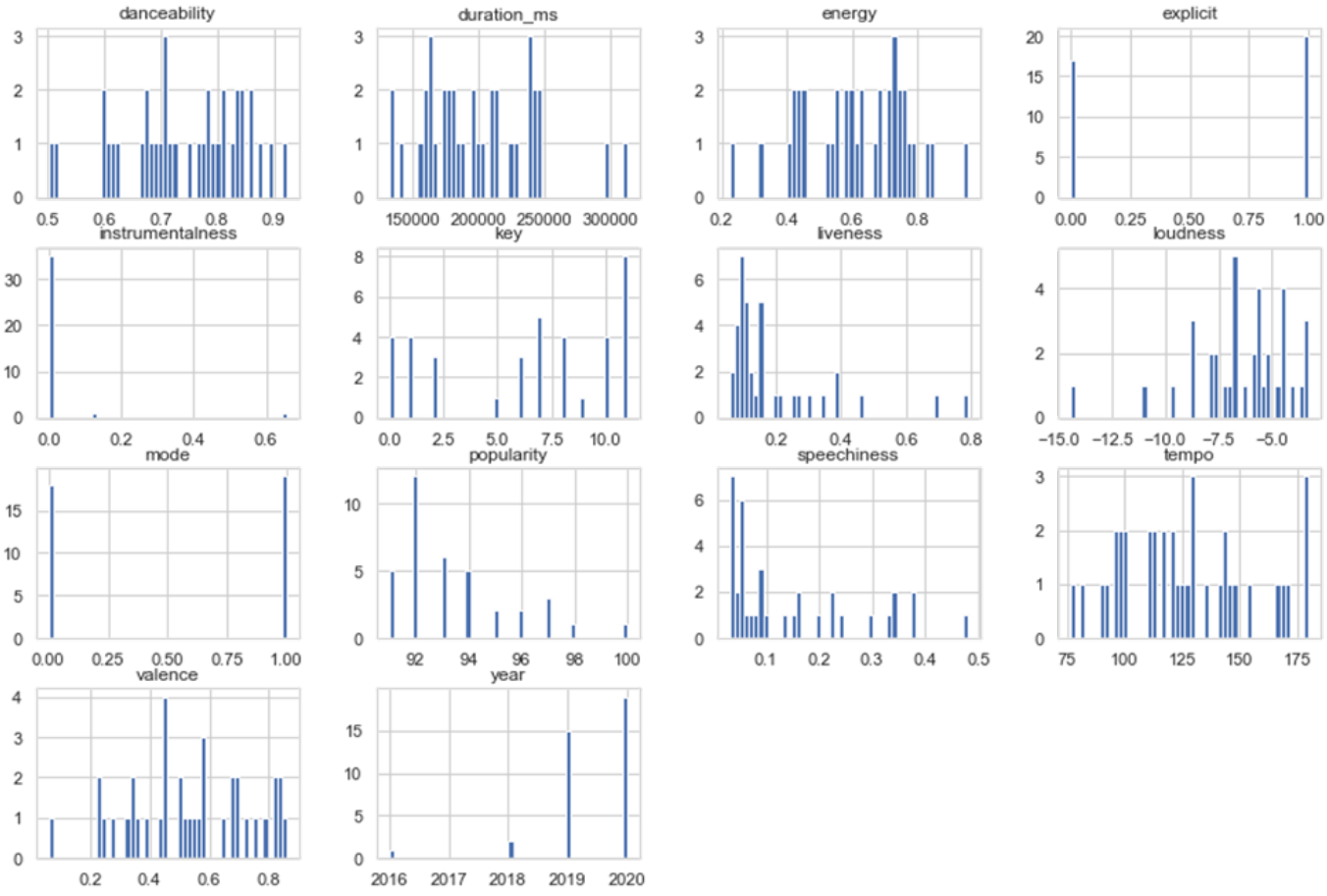

Figure 11. Distribution of audio features after the year 2015 (Millennials)

From the correlation matrix in Figure 2, the popularity and year appear to be highly correlated. The energy, loudness, popularity, and year appear to not correlate with the acousticness, whereas the energy is positively correlated with the loudness. Further to this, Figure 3 shows the details of the correlation matrix. Energy and acousticness appear to have the least correlation of -0.77, and this is followed by acousticness and year having a correlation of -0.64. Each audio feature seems to be self correlated on the matrix with a unit value. However, the highest pair correlation observed is between year with popularity, followed by energy with loudness.

As shown in Figure 4 on audio features correlation with popularity, energy leads, followed by loudness, whereas acousticness has the worst correlation with popularity. In Table 4, the danceability grows as the year increased from 0.425661 in 1921 to 0.57189 in 1925. The energy was building rapidly from 0.236784 in 1921 to 0.348118 in 1924 and then fell to 0.264373 in 1925. The acousticness grew from 0.895823 in 1921 to 0.975369 in 1924 and then rise to 0.965422 in 1925. Different from these trends, the instrumentalness was seen to be rising and falling across the investigated years. On the changes in audio features over time in years, all tested audio features show random variations. The acousticness appears to be more dynamic, while the explicit shows a relatively stable trend.

The top artists of all time, as shown in Figure 6, show that the Beatles are the most popular with around 20000, whereas the Johnny Cash and Beach Boys take the bottom lead with popularity around 13000. For the most popular oldies, as shown in Figure 7, the Beatles retakes the lead with around 18000, followed by the Rolling Stones around 17500. Bob Marley and the Wailers appear to be the least with popularity around 10500. For the audio distribution before the year 2000, danceability shows a normal distribution, acousticness shows a U-shaped distribution. The energy appears to be rising vigorously to an early peak and then fell steadily. The valence is seen to be almost evenly distributed around 3000. The speechiness increases steadily to a peak at 0.17 and then drops drastically to a record low at 0.8. Last, the instrumentalness initially assumes the 80000 extreme but subsequently falls to 0 afterward.

On the most famous artists (Millenials) after the year 2000, as shown in Figure 9, Taylor Swift is seen to be the most renowned artist in the 21st century with the popularity of over 9800, followed by Eminem with the popularity of 8200, while Kanye West is seen as the least popular artist at the time with around 5200.

On the audio feature distribution after the year 2000 (Millenials) in Figure 10, both danceability and valence show normal distribution. The year had all peak, and the popularity and energy maintained the same trend. Figure 11 shows the distribution of audio features after 2015 for Millenials: only the danceability, duration_ms, energy, tempo, and valence show fairly credible results.

5. CONCLUSION

Statistical analysis of the sonic effects of music from a comprehensive dataset on audio features is presented in this paper. The audio features derived from the Spotify web API endpoint and Spotify (Python module for Spotify web servers) were first curated by a third party and rigorously analyzed. This was done in Jupyter notebooks with Python 3, using third-party packages-Pandas and Numpy as analytical tools. Several audio feature parameters such as duration, key, energy, danceability, liveness, notation, mode, time signature, acousticness, instrumentalness, loudness, speechiness, valence, tempo, id, type, and popularity were examined and analyzed. Results indicate a strong correlation amongst the tested parameters, and the danceability appears to follow a standard distribution curve with slight variations in some scenarios. The insights provided by this data would provoke further research in the field of sound and speech synthesis, audio technology, music production, and sound and audio distribution. Future work would focus on developing an efficient audio feature analysis tool to reduce the impact of interference on useful signals.

ACKNOWLEDGMENT

The authors thank Spotify for well-documented access to their API and Yamac Eren Ay to gather useful data and make it available on Kaggle. Agbotiname L. Imoize is partly supported by the Nigerian Petroleum Technology Development Fund (PTDF) and the German Academic Exchange Service (DAAD) through the Nigerian-German Postgraduate Program under Grant 57473408.

REFERENCES

[1] J. Amaegbe and P. Omuku, “Sonic Effects of Indigenous Percussive Musical Instruments in Choral Music Performances : The Case of Kings Choral Voices of Port Harcourt,” *J. Niger. Music Educ.*, no. 9, pp. 180–189, 2017.

[2] T. Wilmering, D. Moffat, A. Milo, and M. B. Sandler, “A history of audio effects,” *Appl. Sci.*, vol. 10, no. 791, pp. 1–27, 2020, doi: 10.3390/app10030791.

[3] P. C. Terry, C. I. Karageorghis, M. L. Curran, O. V. Martin, and R. L. Parsons-Smith, “Effects of Music in Exercise and Sport: A Meta-Analytic Review,” *Psychol. Bull.*, vol. 146, no. 2, pp. 91–117, 2019, doi: 10.1037/bul0000216.

[4] M. Anglada-tort, S. Keller, J. Steffens, and D. Müllensiefen, “The Impact of Source Effects on the Evaluation of Music for Advertising Are there Differences in How Advertising Professionals and Consumers Judge Music?,” *J. Advert. Res.*, no. July, pp. 1–15, 2020, doi: 10.2501/JAR-2020-016.

[5] A. Zelechowska, V. E. Gonzalez-Sanchez, B. Laeng, and A. R. Jensenius, “Headphones or Speakers? An Exploratory Study of Their Effects on Spontaneous Body Movement to Rhythmic Music,” *Front. Psychol.*, vol. 11, no. 698, pp. 1–19, 2020, doi: 10.3389/fpsyg.2020.00698.

[6] T. Theorell and E. Bojner Horwitz, “Emotional Effects of Live and Recorded Music in Various Audiences and Listening Situations,” *Medicines*, vol. 6, no. 16, pp. 1–12, 2019, doi: 10.3390/medicines6010016.

[7] G. Tzanetakis and P. Cook, “MARSYAS: A framework for audio analysis,” *Organised Sound*, vol. 4, no. 3, pp. 169–175, 2000, doi: 10.1017/S1355771800003071.

[8] T. Zhang and C. C. Jay Kuo, “Audio content analysis

for online audiovisual data segmentation and classification," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 441–457, 2001, doi: 10.1109/89.917689.

[9] L. Lu, H. J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 504–516, 2002, doi: 10.1109/TSA.2002.804546.

[10] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio Analysis for Surveilance Applications," in *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 158–161.

[11] B. De Man, B. Leonard, R. King, and J. D. Reiss, "An analysis and evaluation of audio features for multitrack music mixtures," *Proc. 15th Int. Soc. Music Inf. Retr. Conf. ISMIR 2014*, no. ISMIR, pp. 134–142, 2014.

[12] A. Jamdar, J. Abraham, K. Khanna, and R. Dubey, "Emotion Analysis of Songs Based on Lyrical and Audio Features," *Int. J. Artif. Intell. Appl.*, vol. 6, no. 3, pp. 35–50, 2015, doi: 10.5121/ijaia.2015.6304.

[13] T. Giannakopoulos, "PyAudioAnalysis: An open-source python library for audio signal analysis," *PLoS One*, vol. 10, no. 12, pp. 1–17, 2015, doi: 10.1371/journal.pone.0144610.

[14] N. Takahashi, M. Gygli, and L. van Gool, "AENet: Learning deep audio features for video analysis," *IEEE Trans. Multimed.*, vol. 20, no. 3, pp. 513–524, 2018, doi: 10.1109/TMM.2017.2751969.

[15] G. Tzanetakis and P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–301, 2002.

[16] Theimer, Vatolkin, and Eronen, "Definitions of Audio Features for Music Content Description," ... *Rep. TR08-2- ...*, no. Ls 11, 2008.

[17] R. Panda, R. M. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Trans. Affect. Comput.*, pp. 1–14, 2018, doi: 10.1109/TAFFC.2018.2820691.

[18] D.-C. M. L. N. Hanna P., B.-P. J, P. Hanna, N. Louis, M. Desainte-Catherine, and J. Benois-Pineau, "Audio features for noisy sound segmentation," vol. 1, pp. 120–124, 2004.

[19] Z. Cataltepe, Y. Yaslan, and A. Sonmez, "Music genre classification using MIDI and audio features," *EURASIP J. Adv. Signal Process.*, vol. 2007, 2007, doi: 10.1155/2007/36409.

[20] D. Mitrović, M. Zeppelzauer, and C. Breiteneder, "Features for Content-Based Audio Retrieval," vol. 78, pp. 71–150, 2010, doi: 10.1016/s0065-2458(10)78003-7.

[21] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 2, pp. 467–476, 2008, doi: 10.1109/TASL.2007.913750.

[22] J. T. Foote, "Content-Based of Music and Audio," 1997, [Online]. Available: http://www.music.mcgill.ca/~ich/classes/mumt611_05/Query Retrieval/footespie97.pdf.

[23] F. Mörchen, A. Ultsch, M. Thies, and I. Löhken, "Modeling timbre distance with temporal statistics from polyphonic music," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 81–90, 2006, doi: 10.1109/TSA.2005.860352.

[24] R. Neumayer and A. Rauber, "Integration of text and audio features for genre classification in music information retrieval," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4425 LNCS, pp. 724–727, 2007, doi: 10.1007/978-3-540-71496-5_78.

[25] D. Tingle, Y. E. Kim, and D. Turnbull, "Exploring automatic music annotation with 'acoustically-objective' tags," *MIR 2010 - Proc. 2010 ACM SIGMM Int. Conf. Multimed. Inf. Retr.*, pp. 55–61, 2010, doi: 10.1145/1743384.1743400.

[26] S. Shuvaev, H. Giaffar, and A. A. Koulakov, "Representations of Sound in Deep Learning of Audio Features from Music," 2017, [Online]. Available: http://arxiv.org/abs/1712.02898.

[27] M. Müller and S. Ewert, "Towards Timbre-Invariant Audio Features for Harmony-Based Music," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 18, no. 3, pp. 649–662, 2010.

[28] L. Mion and G. De Poli, "Score-independent audio features for description of music expression," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 2, pp. 458–466, 2008, doi: 10.1109/TASL.2007.913743.