IMPROVING INTERPRETABLE GENRE RECOGNITION WITH AUDIO FEATURE STATISTICS BASED ON ZYGONIC THEORY

Igor VATOLKIN (igor.vatolkin@udo.edu) (0000-0002-9454-9402)

Department of Computer Science, **TU Dortmund**, Dortmund, Germany

ABSTRACT

Automatic music genre recognition helps to organise music collections and discover new music pieces. However, state-of-the art classification models often lack of comprehensibility for listeners or musicologists, either because they are based on low-level audio descriptors or complex neural network architectures which operate as black boxes. In this work, we propose novel and interpretable statistics derived from zygonic theory and estimated for discretised audio features. These statistics describe repetitive or similar patterns with respect to meaningful musical properties like energy, harmony, instrumentation, tempo, and rhythm. Compared to the baseline from previous work, which builds models with comprehensible semantic features, we show that for most of tested genres and styles the classification performance increases when semantic features are combined with statistics based on zygonic theory, and the most relevant dimensions are stored after feature selection.

1. INTRODUCTION

Automatic recognition of music genres helps to organise large music collections and recommend new music tracks. This application is particularly close to listeners. However, the decisions made by algorithms are seldom explainable and the prediction models are often not interpretable. For instance, when low-level audio signal descriptors or deep neural networks are involved in genre prediction, it is not easy to describe comprehensible and semantic properties of a genre or simply understand why a music piece is recommended. This can be achieved, when classification models take into account music properties with relation to music theory: instrumentation, harmony, melody, tempo, rhythm, dynamics, etc.

In this work, we have created a set of new audio properties which are based on the zygonic theory proposed by Ockelford [1, 2]. Zygons measure similarity relationships between time series of music properties and were introduced for groups of notes in the score. We have transferred this approach to discretised audio features and proposed statistics which can be used for music genre recognition. After the extension of the large set of semantically interpretable properties from our previous work with these new features, we could observe the reduction of classification errors for almost all tested categories. As an outcome of feature selection applied for complete feature sets, it was possible to identify the most relevant semantic features which contributed to the best classification models with respect to two evaluation criteria: as low classification error as possible and as small feature set as possible.

In Section 2, we start with a brief overview of related work on genre recognition. Section 3 presents the implementation of zygons for audio features and the statistics which were estimated after zygon extraction. Section 4 describes the setup of experiments. The results are discussed in Section 5. We conclude with a summary of main outcomes of the study and ideas for future work.

2. RELATED WORK

In contrast to music genre recognition based on symbolic data like MIDIs, which allow for a rather simple extraction of interpretable music properties [3, 4], the classification models applied for audio signals are typically based on less comprehensible features which were manually engineered by audio signal experts in earlier works or estimated by deep neural networks in more recent studies. [5] formulated recognition of genres in audio signals as automatic classification task based on audio features which described timbre, rhythm, and pitch. A decade later, Sturm already mentioned several hundreds of related studies [6]. More recent studies introduced deep neural networks [7, 8], sometimes integrating further modalities beyond audio [9].

Rather few studies focused on interpretability. Melodic features for genre classification were proposed in [10]. [11] introduced a concept of structural complexity which could be estimated for different features, like rhythm or timbre complexity. This method was applied in [12] for genre recognition with feature sets describing eight musical aspects including chords, harmony, and instruments. Some works focused rather on interpretable classification systems, like fuzzy rules for music genre recognition [13, 14]. In [15], we proposed a set of 566 semantic descriptors with relation to music theory, and formulated music genre recognition as a multi-objective feature selection task which simultaneously minimises the classification error and maximises the share of interpretable semantic features.

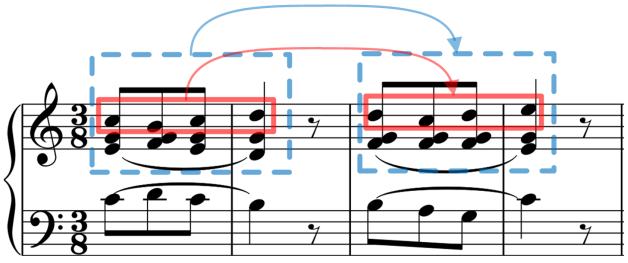

Figure 1. Examples of zygons in the beginning of "Lustig und traurig" by Ludwig van Beethoven.

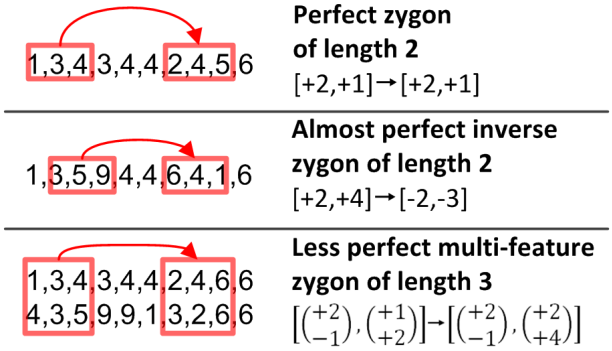

Figure 2. Examples of zygons in discretised feature series.

3. ZYGONS

In the zygonic theory, "it is hypothesized that, at the level of frames, zygonic relationships (or 'zygons') reasonably model a link thought to be constructed cognition between any pairs of perspective values of the same type, where one is deemed to derive from the other" [2]. In other words, zygons characterise similarities between derivations in sequences / series of events which may repeat and be imitated through variation across a music piece. For instance, motifs build *perfect* zygons. Other zygons describe less perfect relationships.

Examples of two zygons are provided in Figure 1, the beginning of "Lustig und traurig" by Ludwig van Beethoven. A perfect zygon with regard to note duration exists between the two chord sequences in dashed blue rectangles. The derivations (changes) in note lengths are $[0, 0, +1/8]$ for both sequencies: three chords are built with eighth notes, the last one with quarter notes. The note sequences in red rectangles show an almost perfect zygon with regard to the note pitch: the halftone changes are $[-1, +1, +2]$ for the first sequence and $[-2, +2, +2]$ for the second one.

The score is not always available for music pieces. Furthermore, audio features may capture relevant characteristics of an artist or genre which can not be extracted from the score: applied digital effects, dynamics, individual deviation of the perfect rhythm, etc. Therefore, we propose the following steps for the identification of similarity relationships between derivations of audio features:

1. Selection of one or several base features which represent some meaningful semantic concept, for instance, timbre, semitone spectrum, or dynamics.

2. Aggregation of original feature values: as many audio features are extracted from very short time frames like 23 ms (512 sample frame after the fast Fourier transform with a sampling rate of 22050 Hz), they should be aggregated for longer time frames closer to typical note durations.

3. Discretisation: similar to discretised pitch height and duration in the score, continuous feature values are mapped to a limited number of discrete values or n-grams. The original values can be assigned to quantiles or histogram bins: e.g., the first of four quartiles describes all features with values which belong to the lowest 25% of all sorted continuous values from the feature series in a music track.

4. Derivation computation: for the analysis of changes in a group of succeeding frames, derivations of original feature series are estimated. Also derivations of second and higher orders may be considered.

Three examples for zygons are visualised in Figure 2: a perfect zygon, an almost perfect inverse zygon, and a less perfect zygon. The last zygon shows how the derivations are estimated for several features or multiple dimensions of an individual feature, e.g., all dimensions of the Mel frequency cepstral coefficients [16] (MFCCs). Thus, we distinguish here between *single-feature* zygons and *multi-feature* zygons.

For an overview of parameters which have an impact on zygon calculation, see Table 3 in Section 4.

4. DESIGN OF EXPERIMENTS

4.1 Datasets and Baseline with Semantic Features

As [15] focuses on interpretability and provides a large set of 566 semantic features which describe very different music properties, we use this setup as a baseline. Table 1 provides an overview of these semantic descriptors.

For a better comparison with [15], we use the same dataset for the experiments. 6 genres and 8 music styles from the Ls11 collection [1] are predicted. The training sets for supervised binary classification tasks are compiled from 10 "positive" and 10 "negative" tracks per category. This is motivated by a realistic application to prevent listener fatigue who selects only a limited number of tracks for the automatic classification system. Because of small training sets and the main focus on interpretability, the classification is done with an ensemble of four traditional methods (decision tree C4.5, naive Bayes, random forest, support vector machine); learning of genres with deep neural networks in this setup would lead to a high risk of overfitting and a poor interpretability.

As the four mentioned classifiers do not always lead to very interpretable models even if they are trained with semantic features (consider very large decision trees), the

[1] https://ls11-www.cs.tu-dortmund.de/rudolph/mi#music_test_database; accessed on 03.11.2021

Feature Group	Examples
Chroma and harmony	Strengths of intervals in chroma DCT-reduced log pitch (CRP) [17], consonance [18], key and its clarity [19]
Chord statistics	Number of different chords in 10 s
Predicted instruments	Share of guitars, wind, strings
Predicted moods	Energetic, sentimental, earnest
Predicted descriptors from [20]	Female or male vocals, melodic range, activation grade
Tempo and rhythm	Estimated onset number per minute, rhythmic clarity [19]
Structural complexity after [11]	Complexity of chords, instruments, chroma, harmony

Table 1. Groups of baseline semantic audio features.

evolutionary multi-objective feature selection, applied for each classification task separately, as in [15], provides a good possibility to identify in a non-deterministic way different feature subsets which are particularly successful for genre prediction models. For ten statistical repetitions, randomly initialised feature sets are evolved for an ensemble of four classifiers, with the goal to simultaneously minimise the balanced classification error and the number of selected features. Then, we can measure relative occurrences of baseline semantic features and statistics of zygons in *non-dominated* fronts of trade-off feature sets which are the best compromise solutions with respect to both criteria. The feature selection is evaluated on a validation set of 120 tracks, and the last population of feature sets is finally evaluated on a test set of 120 tracks with the same genre distribution but other artists and albums than in training and optimisation sets.

Let TP denote the true positives (tracks which belong to the current genre and are predicted as belonging to it), TN the true negatives (tracks which do not belong to the genre and are predicted as not belonging to it), FP the false positives, and FN the false negatives. The balanced classification error m_e is defined as

$$m_e = \frac{1}{2}\left(\frac{FN}{TP+FN} + \frac{FP}{TN+FP}\right) \qquad (1)$$

and is motivated by the situation that the validation and test sets are not balanced in the distribution of "positive" and "negative" tracks.

The second optimisation criterion, the number of selected features, favours smaller feature sets which are more interpretable and also typically lead to more robust classification models compared to models constructed with very large feature sets.

4.2 Zygonic Descriptors

Table 2 lists 39 base audio feature sets carefully selected after preliminary experiments for the estimation of multi-feature zygons. Please note that even if some zygons are identified in low-level audio descriptors like MFCCs, their meaning has a higher interpretability, such as "measured relationships in timbre changes based on MFCCs".

The list of final parameters for the estimation of zygons in this study is provided in Table 3. Some of them were adjusted after preliminary experiments (e.g., the base feature aggregation window length of 250 ms and 500 ms performed better than 1 s for short-framed features, or discretisation with quartiles better than with 8 quantile bins). However, the exhaustive evaluation and statistical analysis of all possible parameter combinations was beyond of scope of this work. Some promising directions are discussed in Section 6.

The analysis window length of 30 s performed better than other lengths. Besides, as we expect certain variations of semantic music properties over music segments like verse or bridge, we have extracted segments with a convolutional neural network (CNN) from [26] trained with SALAMI dataset [27], so that zygons were estimated for these segments and not analysis windows of a fixed length.

For the estimation of a perfection grade, the difference between derivations of discretised features is estimated as follows. Let d be the number of all base feature dimensions and t the length of two sequences of derivations to be compared. Let u_{ij} and v_{ij} denote the derivations of i-th dimension at j-th position of the sequences u resp. v. The difference between both sequences $g(u, v)$ is measured as:

$$g = \sum_{i=1}^{d}\sum_{j=1}^{t} |u_{ij} - v_{ij}|. \qquad (2)$$

The intervals which map g-values to five categories of perfection grades are provided in the last row of Table 3; they were adjusted after some preliminary experiments without a claim to be exhaustive.

After the estimation of zygons, we measure the following two statistics: the number of different zygons and the overall number of sequences contributed to zygon estimation (for each zygon, this number is reduced by 1, as at least two sequences are required to identify a zygon). For example, if the first zygon from Figure 1 $[+2, +1]$ is identified as perfect zygon across 6 sequences in the analysis window, another zygon $[+5, +1]$ across 4 sequences, and no other perfect zygon is found, the number of different perfect zygons is 2 and the overall number of sequences 8.

Thus, the overall number of zygon statistics is 39 (number of base features) $\times 2$ (direct and inverse zygons) $\times 5$ (number of perfection grades) $\times 2$ (two different zygon lengths) $\times 2$ (two analysis window lengths) $\times 2$ (two zygon statistics) $= 3120$. The experiments with zygons based on quartile and histogram discretisation were carried out independently of each other.

Base Features
Energy
• All energy features: low energy, sub-band energy ratio, root mean square, zero-crossing rate
• Root mean square of the time signal
• Zero-crossing rate
Harmony
• A large set of harmonic features: amplitude and tone of maximum strength in chromagram, consonance, fundamental frequency, harmonic change detection function, inharmonicity, key and its clarity, local tuning, major/minor alignment, tonal centroid
• Chord vector [12]
• Chroma DCT-reduced log pitch [17]
• Tonal centroid [19]
• Semitone spectrum [18]
• Strengths of major / minor keys [19]
• Interval strengths from 10 highest semitone peaks
• Interval strengths from the semitone peaks above 3/4 of the maximum peak
Instruments: approximative features [21]
• Dominance ranks of 51 instruments
• Mean similarities to 51 instruments
• Similarities to bowed instruments
• Similarities to drums
• Similarities to key instruments
• Similarities to pianos
• Similarities to stringed instruments
• Similarities to Western strings
• Similarities to Western guitars
• Similarities to woodwind instruments
Instruments: CNN predictions [22]
• All instruments
• Bowed instruments
• Drums
• Key instruments
• Pianos
• Stringed instruments
• Western strings
• Western guitars
• Woodwind instruments
Pitches: approximative features [21]
• All pitches
Rhythm and tempo
• Fluctuation patterns [19]
• Rhythmic clarity [19]
• Sum of correlated components, periodic. peaks [23]
Timbre
• Mel frequency cepstral coefficients [16]
• A large set of timbre features: angles and distances in phase domain, sensory roughness, spectral moments and further spectral properties, tristimulus, etc.
• Angles and distances in phase domain [24]
• Spectral moments [25]
• Tristimulus and normalised energy of harmonic components [25]

Table 2. Base audio features for zygon estimation.

5. RESULTS

Table 4 contains the smallest balanced classification errors estimated for artist- and album-independent test sets for all final outcomes after evolutionary multi-objective feature selection. The baseline with 566 semantic descriptors from [15] is reported in the 2nd column ("Sem"). Columns "ZygQ" and "ZygH" contain values for zygon statistics estimated with quartile discretisation resp. histogram discretisation. Columns "ZygQ+Sem" and "ZygH+Sem" correspond to combined feature sets. Errors lower than for the baseline are highlighted with a bold font, and the smallest error for each genre or style is underlined.

The zygon statistics alone outperform the baseline for only 5 resp. 6 categories—which can be also expected—as this set contains a very large number of various semantic descriptors. However, if we start with combined feature sets, the errors are reduced for 11 of 14 categories, in particular for all but one style (Alternative Pop Rock, histogram-based zygons). When we take both zygon statistics "ZygQ" and "ZygH" into account, only for genres Pop and R'n'B the baseline performs at best. Please note that the genre Pop was ambiguous and compiled from very different styles including Synth Pop, Progressive Rock, and Heavy Metal.

Sometimes, combined feature sets led to higher errors (e.g., for genre Electronic it was better to use only zygon statistics and not to combine them with the baseline). This is explained by a strict evaluation on an independent test set. Using too many features increases an effect of overfitting towards the validation set for the evaluation of selected features, even if this effect is not crucial for tested genres: the baseline test error for Electronic is still significantly higher.

6. CONCLUSIONS AND OUTLOOK

In this work, we have applied the zygonic theory for the search of similarity relationships in audio feature series which describe different musical properties like harmony, instrumentation, or energy. We have proposed statistics based on zygons which can be used for music genre recognition. The combination with another large set of semantic audio descriptors and the application of multi-objective feature selection helped to improve the classification performance for 12 of 14 genres and styles.

The success of this initial study is especially promising, as we could investigate only a very limited range of different possible parameter settings for zygon extraction, from the base audio features to the discretisation method, the length of aggregation window, the length of sequences to analyse, etc. More exhaustive studies are planned for future. Also, further statistics of extracted zygons can be constructed, for example occurrences of specific zygons which may characterise particular genres.

For a better reliability of our observations, we plan to extend experiments to larger datasets. However, this remains a very challenging task, because the extraction of audio features and zygons requires long times and for many state-of-the-art datasets the audio is either not available or only

Parameter	Setting
Base feature	39 sets listed in Table 2
Base feature aggregation window length	500 ms for approximative features after [21], 1500 ms for fluctuation patterns, 3000 ms for other tempo/rhythm features, 250 ms for all remaining features
Discretisation method	Quartiles, histograms with 10 bins
Analysis window length	30 s with 15 s step size, segments extracted after [26]
Dimensionality	Multi-feature zygons
Zygon length	2, 3
Zygon type	Direct, inverse
Perfection grades	Perfect, almost perfect, more or less perfect, less perfect, imperfect
Boundary for estimation of perfection grades	$g \in [0, \lceil \frac{d}{2} \rceil]$ for perfect zygons, $g \in [\lceil \frac{d}{2} \rceil + 1, d + 1]$ for almost perfect zygons, $g \in [d + 2, \lceil \frac{3}{2} d \rceil]$ for more or less perfect zygons, $g \in [\lceil \frac{3}{2} d \rceil + 1, 2d + 2]$ for less perfect zygons, and $g \in [2d + 3, \lceil \frac{5}{2} d \rceil + 2]$ for imperfect zygons

Table 3. Parameters for zygon calculation.

Genre/ style	Sem	ZygQ	ZygQ +Sem	ZygH	ZygH +Sem
Classic	0.0276	0.0333	**0.0238**	0.0857	**0.0048**
Electr	0.1610	**0.0619**	**0.0801**	**0.0762**	**0.0952**
Jazz	0.1400	0.1429	0.1571	0.1524	**0.1381**
Pop	0.1575	0.3111	0.2244	0.2800	0.2244
Rap	0.0642	**0.0286**	**0.0238**	**0.0333**	**0.0143**
R'n'B	0.1458	0.1619	0.1857	0.1571	0.2000
Adult	0.2417	**0.2273**	**0.2409**	0.2773	**0.2318**
AlbR	0.2316	0.2364	**0.1545**	0.2909	**0.1773**
AltPR	0.2251	0.2508	**0.2192**	**0.1727**	0.2327
Club	0.1760	0.2601	**0.1395**	**0.1489**	**0.1415**
Heavy	0.1213	0.1562	**0.0848**	0.1786	**0.0982**
Prog	0.2309	**0.2212**	**0.2212**	**0.2212**	**0.2168**
Soft	0.1862	0.2312	**0.1486**	0.2177	**0.1847**
Urban	0.2061	**0.1739**	**0.2000**	**0.1957**	**0.2000**

Table 4. Smallest m_e for the test sets after feature selection. Sem: semantic baseline features; ZygQ: zygon statistics based on quartile discretisation; ZygH: based on histogram discretisation; +Sem: zygon statistics together with semantic baseline. Abbreviated genres/styles: Electr: Electronic; Adult: Adult Contemporary; AlbR: Album Rock; AltPR: Alternative Pop Rock; Club: Club Dance; Heavy: Heavy Metal; Prog: Prog Rock; Soft: Soft Rock.

short excerpts exist (like 30 s previews for the Million Song Dataset [28]). Another promising application scenario is the analysis of individual composing styles and their development over time.

Acknowledgments

This work was funded by the German Research Foundation (DFG), project 336599081 "Evolutionary optimisation for interpretable music segmentation and music categorisation based on discretised semantic metafeatures". The experiments were carried out on the Linux HPC cluster at TU Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the German Research Foundation (DFG) as project 271512359.

7. REFERENCES

[1] A. Ockelford, "The role of repetition in perceived musical structures," in *Representing Musical Structure*. London: Academic Press, 1991, pp. 129–160.

[2] ——, "On similarity, derivation and the cognition of musical structure," *Psychology of Music*, vol. 32, no. 1, pp. 23–74, 2004.

[3] C. McKay, J. Cumming, and I. Fujinaga, "jSymbolic 2.2: Extracting features from symbolic music for use in musicological and MIR research," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, E. Gómez, X. Hu, E. Humphrey, and E. Benetos, Eds., 2018, pp. 348–354.

[4] E. Zheng, M. Moh, and T.-S. Moh, "Music genre classification: A n-gram based musicological approach," in *Proceedings of the 7th IEEE International Advance Computing Conference (IACC)*, 2017, pp. 671–677.

[5] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[6] B. L. Sturm, "A survey of evaluation in music genre recognition," in *Proceedings of the 10th International Workshop on Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation (AMR)*, 2012, pp. 29–66.

[7] A. R. Rajanna, K. Aryafar, A. Shokoufandeh, and R. Ptucha, "Deep neural networks: A case study for music genre classification," in *Proceedings of the 14th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2015, pp. 655–660.

[8] L. Nanni, Y. M. G. Costa, R. L. Aguiar, C. N. S. Jr., and S. Brahnam, "Ensemble of deep learning, visual and acoustic features for music genre classification," *Journal of New Music Research*, vol. 47, no. 4, pp. 383–397, 2018.

[9] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, pp. 4–21, 2018.

[10] J. Salamon, B. M. M. Rocha, and E. Gómez, "Musical genre classification using melody features extracted from polyphonic music signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 81–84.

[11] M. Mauch and M. Levy, "Structural change on multiple time scales as a correlate of musical complexity," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, A. Klapuri and C. Leider, Eds. University of Miami, 2011, pp. 489–494.

[12] P. Ginsel, I. Vatolkin, and G. Rudolph, "Analysis of structural complexity features for music genre recognition," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2020, pp. 1–8.

[13] F. Fernández and F. Chávez, "Fuzzy rule based system ensemble for music genre classification," in *Proceedings of the 1st International Conference on Evolutionary and Biologically Inspired Music, Sound, Art and Design (EvoMUSART)*, ser. Lecture Notes in Computer Science, P. Machado, J. Romero, and A. Carballal, Eds., vol. 7247. Springer, 2012, pp. 84–95.

[14] F. Heerde, I. Vatolkin, and G. Rudolph, "Comparing fuzzy rule based approaches for music genre classification," in *Proceedingsof the 9th International Conference on Artificial Intelligence in Music, Sound, Art and Design (EvoMUSART 2020)*, ser. Lecture Notes in Computer Science, J. Romero, A. Ekárt, T. Martins, and J. Correia, Eds., vol. 12103. Springer, 2020, pp. 35–48.

[15] I. Vatolkin, G. Rudolph, and C. Weihs, "Interpretability of music classification as a criterion for evolutionary multi-objective feature selection," in *Proceedings of the 4th International Conference on Evolutionary and Biologically Inspired Music, Sound, Art and Design (EvoMUSART)*, ser. Lecture Notes in Computer Science, C. G. Johnson, A. Carballal, and J. Correia, Eds., vol. 9027. Springer, 2015, pp. 236–248.

[16] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River: Prentice Hall, 1993.

[17] M. Müller and S. Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011, pp. 215–220.

[18] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, J. S. Downie and R. C. Veltkamp, Eds., 2010, pp. 135–140.

[19] O. Lartillot and P. Toiviainen, "MIR in Matlab (II): A toolbox for musical feature extraction from audio," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, 2007, pp. 127–130.

[20] G. Rötter, I. Vatolkin, and C. Weihs, "Computational prediction of high-level descriptors of music personal categories," in *Algorithms from and for Nature and Life - Classification and Data Analysis*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, B. Lausen, D. V. den Poel, and A. Ultsch, Eds. Springer, 2013, pp. 529–537.

[21] I. Vatolkin, "Evolutionary approximation of instrumental texture in polyphonic audio recordings," in *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2020, pp. 1–8.

[22] I. Vatolkin, B. Adrian, and J. Kuzmic, "A fusion of deep and shallow learning to predict genres based on instrument and timbre features," in *Proceedings of the 10th International Conference on Artificial Intelligence in Music, Sound, Art and Design (EvoMUSART)*, ser. Lecture Notes in Computer Science, J. Romero, T. Martins, and N. Rodríguez-Fernández, Eds., vol. 12693. Springer, 2021, pp. 313–326.

[23] D. McEnnis, C. McKay, I. Fujinaga, and P. Depalle, "jAudio: An feature extraction library," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, 2005, pp. 600–603.

[24] I. Mierswa and K. Morik, "Automatic feature extraction for classifying audio data," *Machine Learning Journal*, vol. 58, no. 2-3, pp. 127–149, 2005.

[25] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," IRCAM, Tech. Rep., 2004.

[26] T. Grill and J. Schlüter, "Music boundary detection using neural networks on combined features and two-level annotations," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, M. Müller and F. Wiering, Eds., 2015, pp. 531–537.

[27] J. B. Smith, J. A. Burgoyne, I. Fujinaga, D. D. Roure, and J. S. Downie, "Design and creation of a large-scale database of structural annotations," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, A. Klapuri and C. Leider, Eds. University of Miami, 2011, pp. 555–560.

[28] T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, A. Klapuri and C. Leider, Eds. University of Miami, 2011, pp. 591–596.