3.1 TIV Complexity Features

Our first feature group, denoted as **Q**, captures aspects of tonal complexity and consonance, somehow mirroring the features presented in Section 2.2.

3.1.1 Distance Between Audio Segments

Distances between TIVs relate to their perceived similarity such that small distances equate to perceptually similar sonorities. To this end, Euclidean and cosine distance metrics between TIVs, which measure different perceptual characteristics of the signal, have been considered in the literature [14, 27].

The cosine distance θ measures the phase difference between two TIVs, capturing pitch class distances that roughly correspond to voice leading parsimony. Smaller values indicate a higher number of shared pitch classes.

The Euclidean distance metric d captures the phase φ_k and magnitude $\|T_k\|$ differences between TIVs. In addition to measuring the number of shared pitch classes between TIVs, it also accounts for shared interval content.

To exploit the musical properties of these two distance metrics, we propose new features that measure the perceptual distance between consecutive musical audio segments. Given the TIV $\mathbf{T}_n = (T_{0,n}, \ldots, T_{5,n})$ of the n-th segment, the Euclidean (d) and cosine (θ) distances between two consecutive segments is defined as:

$$\begin{aligned} Q_{\text{EucDist}} &= s_n^{\text{euc}} = d\{\mathbf{T}_n, \mathbf{T}_{n+1}\}, \\ Q_{\text{CosDist}} &= s_n^{\text{cos}} = \theta\{\mathbf{T}_n, \mathbf{T}_{n+1}\} \end{aligned} \tag{3}$$

The metrics can be applied at multiple hierarchies (or time scales) by adopting musical audio segments of different sizes. While short segments capture chord changes, larger segments can capture the overall tonal structure (e.g., key modulations).

3.1.2 Tonal Dispersion

The tonal dispersion of the n-th audio segment n is defined as the distance of its respective TIV to the tonal center $\overline{\mathbf{T}}$:

$$\begin{aligned} Q_{\text{EucTonalDisp}} &= u_n^{\text{euc}} = d\{\mathbf{T}_n, \overline{\mathbf{T}}\}, \\ Q_{\text{CosTonalDisp}} &= u_n^{\text{cos}} = \theta\{\mathbf{T}_n, \overline{\mathbf{T}}\} \end{aligned} \tag{4}$$

The tonal center of a piece corresponds to the TIV of the pitch-class distribution averaged across its entire duration. The tonal dispersion indicates how much the harmonic content of a given segment deviates from the tonal center $\overline{\mathbf{T}}$ and can indicate the degree of modulations across the piece or the segments with non-diatonic pitch classes (i.e., further away from the tonal center). This feature can help discriminate later classical musical periods that typically have larger tonal dispersion [28]. To compute the distance of a given segment n from the tonal center, we adopt both the Euclidean and cosine distance metrics to capture the different harmonic aspects mentioned in Section 3.1.1.

3.1.3 TIV Entropy

Related literature has been adopting Shannon entropy to capture the harmonic complexity within musical style [11, 29]. Amiot [30] recently proposed the computation of harmonic complexity from symbolic pitch-class distributions as the Shannon entropy of its Fourier coefficients. Following [30], we propose a TIV entropy feature to capture harmonic complexity from musical audio. The TIV entropy is high for random pitch-class distributions and low for pitch-class distributions that exhibit some degree of organization (i.e., periodicity or sparseness). We define the TIV entropy feature as the entropy of the normalized coefficient magnitudes of a TIV $\mathbf{T}$:

$$Q_{\text{TIVEntropy}} = \sum_{k=1}^{6} -p_k \log p_k, \quad p_k = \frac{\|T_k\|}{\sum_{j=1}^{6} \|T_j\|} \tag{5}$$

While entropy has been shown to capture different stylistic attributes depending on the time scale under analysis [29], we demonstrate in Table 2 the TIV entropy of various pitch-class sets. Lower TIV entropy (i.e., denoting greater organization) includes fairly common pitch-class distributions such as diatonic scales, triads, and tetrads. Higher TIV entropy denotes musical structures with less common adoption in the tonal music lexica, such as the chromatic set and its many subsets.

Pitch class set	TIV Entropy
Octatonic scale	0.3551
Diminished seventh chord	0.3552
Whole tone scale	0.5268
Diatonic scale	1.2444
Pentatonic scale	1.2972
Harmonic minor scale	1.4927
Minor seventh chord	1.5542
Single note	1.6767
Chromatic scale	1.6906

Table 2. TIV entropy of several pitch-class sets. For details on the elements of each set please refer to [30].

3.2 Harmonic Rhythm Features

Harmonic rhythm is the rate at which the chords change in a musical piece. It can provide a prominent indicator to distinguish style periods. For example, a fast harmonic rhythm is a characteristic of the Baroque period [31] in comparison with the large-scale structures of a Romantic symphony. To this end, we advance new features (denoted as **R**) that study the rate and magnitude of harmonic changes based on the Harmonic Change Detection Function (HCDF) proposed in [32]. The value ξ_n of the HCDF at frame n is defined as the rate of change between the TIVs $\mathbf{T}_{n-1}$ and $\mathbf{T}_{n+1}$ after Gaussian smoothing and is computed using the Euclidean distance.

Let us now consider Ξ_m as the set of frame indices of the peaks of the HCDF, considering all local maxima as peaks. We define the inter-peak interval feature as the difference between the frame numbers of consecutive peaks of the HCDF:

$$R_{\text{HCDFPeakInterval}} = \Delta_m = \Xi_{m+1} - \Xi_m \tag{6}$$

As an additional feature, we consider the magnitude of the peaks, $R_{\text{HCDFPeakMag}}$, given by ξ_Y, with $Y = \Xi_m$, which captures the degree of the harmonic changes.

3.3 Temporal Resolution of Tonal Interval Vectors

We detail two segmentation strategies for the feature extraction process: multiple fixed-size segment resolutions (Section 3.3.1) and variable-size segments resulting from the analysis of harmonic changes (Section 3.3.2).

3.3.1 Fixed-size Segmentation with Multiple Resolutions

Fixed-size segments with equal duration are adopted at multiple time scales or resolutions to capture the piece's different harmonic or tonal dimensions. Following [10–12], we consider four time resolutions in our work: 100ms, 500ms, 10s, and global (i.e., the entire piece or excerpt under analysis). Smaller time scales (e.g., 100ms and 500ms) capture finer musical elements such as individual notes, intervals, and chords. Larger time scales (e.g., 10s and global) capture higher harmonic structures at the level of structural parts or the overall piece, enhancing the harmonic changes in the large-scale tonal structure of the composition (such as key and modulations) and the harmonic trajectories of the horizontal structure (such as chord progressions).

3.3.2 Harmonic Structural Segmentation

We define a context-sensitive segmentation at the peaks of the HCDF (see Section 3.2). This strategy considers prominent harmonic changes as segment boundaries, therefore producing segments of varying duration according to the specific harmonic changes in the audio content. The use of context-sensitive segmentation in the feature extraction process of musical analysis systems has been shown to improve their accuracy compared to fixed segmentation approaches, which are less aware of musical structure [33, 34].

4. EXPERIMENTS AND RESULTS

To assess the degree to which the harmonic information conveyed by TIV harmonic audio features in Section 3 discriminate musical style, we evaluate the proposed features in two classical music style classification tasks from musical audio: historical periods (e.g., Baroque or Romantic) and composers. Furthermore, we compare the accuracy of the above features to the state-of-the-art template-based and complexity features described in Sections 2.1 and 2.2. To study the accuracy of different groups of features, we consider the two following sets: (template-based $\mathbf{F}$ and tonal complexity $\mathbf{G}$) and (TIV basic $\mathbf{B}$, TIV complexity $\mathbf{Q}$, harmonic rhythm $\mathbf{R}$). Finally, we inspect the impact of different context-sensitive vs. fixed-window segmentation strategies.

4.1 Experimental Procedure

For our experiments, we consider the *Cross-Era* [1] and *Cross-Composer* [2] datasets, which include 1600 and 1100 pieces, respectively. In detail, the *Cross-Era* dataset has 400 pieces per classical style period and features the following four periods: Baroque, Classical, Romantic, and Modern. The *Cross-Composer* dataset includes 100 pieces for each of the 11 featured classical music composers across all style periods. These datasets are further divided into the subsets presented in Table 3. The datasets provide pre-extracted NNLS chroma features at a resolution of 100ms (10Hz) for each piece.

Dataset	No. Classes (Z)	Items per Class
Cross-Era-Full	4	400
Cross-Era-Piano	4	200
Cross-Era-Orchestra	4	200
Cross-Comp-11	11	100
Cross-Comp-5	5	100

Table 3. Balanced subsets obtained from the *Cross-Era* and *Cross-Composer* datasets [12].

Based on [12], we employ the following classification procedure. Using the subsets listed in Table 3, we compute the features and calculate their piece-wise mean (μ) and standard deviation (σ). Then, we perform a stratified split into three cross-validation (CV) folds, one for testing and two others for training the model. Next, we use the two training folds to compute a Linear Discriminant Analysis (LDA) matrix to reduce the dimensionality of all three folds to $L = Z - 1$ with Z being the number of classes per task. We then perform a five-fold grid search on the two training folds to train an SVM classifier while optimizing its hyperparameters. We conduct this procedure three times, with each fold serving for testing once. To evaluate the robustness of the model concerning the random distribution into folds, we repeat the procedure ten times with re-initialized folds to calculate the accuracy deviation between the different classes (i.e., inter-class deviation).

To counteract problems stemming from adopting tracks from the same CD recording on the training and test folds (known as the album effect [35]), we take additional measures to prevent the model from adapting to the acoustic conditions of specific recordings. To this end, inspired by [11], we apply a composer filter (for period classification) or a performer filter (for composer classification) during the CV procedure that forces pieces from the same composer/performer to be placed within the same fold, thus avoiding overfitting while making the task more challenging and closer to real-world application scenarios.

4.2 Influence of Different Types of Segmentation

We analyze the impact of different audio segmentation strategies on the classification of classical music periods. Table 4 displays the classification accuracy for the TIV Basic ($\mathbf{B}$), TIV Complexity ($\mathbf{Q}$), and Harmonic Rhythm ($\mathbf{R}$) feature groups, as well as the combination of all the above features (through concatenation), on the *Cross-Era* dataset. As segmentation strategies, we consider fixed-size segmentation with a single resolution of 100ms (FS), fixed-size segmentation with multiple resolutions (MR)—100ms, 500ms, 10s, and global—and harmonic structural

[1] https://www.audiolabs-erlangen.de/resources/MIR/cross-era
[2] https://www.audiolabs-erlangen.de/resources/MIR/cross-comp

	FS	MR	HS	MR + HS
Cross-Era-Piano				
TIV Basic (B)	57.54%	**66.84%**	51.65%	65.64%
TIV Complexity (Q)	56.03%	**57.99%**	56.07%	57.67%
Harm. Rhythm (R)	-	-	21.26%	-
Combined	62.61%	**65.47%**	48.97%	63.80%
Cross-Era-Orchestra				
TIV Basic (B)	66.84%	74.80%	64.22%	**75.34%**
TIV Complexity (Q)	67.35%	69.87%	64.80%	**70.00%**
Harm. Rhythm (R)	-	-	28.31%	-
Combined	73.80%	77.19%	68.89%	**77.78%**
Cross-Era-Full				
TIV Basic (B)	60.41%	**70.13%**	57.24%	70.08%
TIV Complexity (Q)	55.97%	62.18%	57.35%	**62.86%**
Harm. Rhythm (R)	-	-	21.65%	-
Combined	64.94%	**71.63%**	62.30%	70.99%

Table 4. Classification accuracy for different types of segmentation: fixed-size (FS), fixed-window with multiple temporal resolutions (MR), harmonic structural segmentation (HS), and a combination of the last two approaches.

segmentation (HS). Additionally, we consider combining the two approaches presented in Section 3.3 (MR + HS).

For the *Cross-Era-Piano* and *Cross-Era-Full* datasets, the MR strategy shows slightly higher accuracy than other segmentation strategies. For other scenarios such as the *Cross-Era-Orchestra* dataset, combining the MR and HS approaches leads to similar or slightly better results. Overall, the MR strategy obtains the highest accuracy values in most cases and is less computationally expensive than the HS approach. Therefore, we opt for the MR strategy in all subsequent experiments, using the HS strategy only for the harmonic rhythm feature group $\mathbf{R}$.

4.3 Style Period and Composer Classification

Using the MR strategy, we perform style period and composer classification experiments on a larger set of feature combinations and show the results in Table 5. We first compare the template-based ($\mathbf{F}$) and TIV basic ($\mathbf{B}$) feature groups, which capture similar harmonic aspects. Indeed, we observe that these groups perform similarly on most datasets, with accuracy differences of 2.98% and 4.74% on the *Cross-Era-Piano* and *Cross-Comp-5* datasets, respectively. The tonal complexity ($\mathbf{G}$) and TIV complexity ($\mathbf{Q}$) groups show a greater performance accuracy difference in the *Cross-Era-Piano* dataset (7.85%) and a smaller difference in the *Cross-Era-Full* dataset (3.33%). Concerning the two largest datasets (*Cross-Era-Full* and *Cross-Comp-11*), the best feature groups result from combining TIV features with those proposed in [10, 11], achieving 74.04% and 38.25% accuracy, which correspond to an improvement of 2.88% and 4.74%, respectively, compared to using only the features proposed in previous work.

From these findings, we draw two main conclusions. First, conceptually similar feature groups lead to quite similar classification accuracies—an encouraging finding, which proves that our approach is valid (a kind of "sanity check" against [10, 11]) and indicates that the features exhibit the intended mid-level semantic properties (feature groups describing related harmonic properties behave similarly for classification). Second, the novel features at least partially capture complementary information such that their combination improves upon individual groups. Overall, we do not reach the state-of-the-art results on the *Cross-Era-Orchestra* and *Cross-Era-Full* subsets, which were obtained in [13] using chord transition features in combination with template-based and complexity features, surpassing our results by 6.01% and 4.16%. However, since TIV-based features improve upon the ones of [10,11], we hypothesize that, in combination with the chord transition features of [13], they might be able to reach even better classification accuracies.

4.4 Harmonic Features Correlation

We adopt hierarchical clustering to assess the degree of information (and redundancy) of harmonic audio features. Figure 1 shows the hierarchical clustering of all harmonic features detailed in Sections 2.2, 2.1, and 3, which we adopt in our experiments. Features are computed at a fixed-size 100ms resolution with average and standard deviation statistics.

Template-based ($\mathbf{F}$) and TIV basic ($\mathbf{B}$) features appear strongly correlated, often ending up in the same cluster. This suggests they may be capturing similar musical properties. For example, the two top-most features $F_{\mathrm{IC5},\sigma}$ and $B_{\mathrm{Diatonicity},\sigma}$ are direct neighbours and are both based on perfect-fifth relationships. Tonal Complexity ($\mathbf{G}$) and TIV Complexity ($\mathbf{Q}$) features also appear correlated, albeit to a slightly lower degree. For example, the green cluster is mainly composed of features from these groups. On the other hand, the red cluster contains several TIV Complexity features and none from the Tonal Complexity group, suggesting that those, in particular, describe different harmonic characteristics. The adopted global statistics, mean and standard deviation, tend to be grouped under the same cluster, suggesting to capture complementary information.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a novel set of mid-level and perceptually-inspired harmonic audio features based on TIVs, which provide indicators for musical dissonance, chromaticity, dyadicity, triadicity, diminished quality, diatonicity, and whole-toneness, as well as quantify perceptual relations of short- and long-term harmonic structures, tonal dispersion, harmonic changes, and complexity. Features are computed using context-sensitive harmonic structure segmentation and a fixed-size segmentation with multiple temporal resolutions.

Proposed TIV harmonic features were assessed in two classical music style period and composer classification tasks using five different datasets. The novel TIV harmonic features have been compared to state-of-the-art harmonic features proposed in [10–12] and showed similar results, slightly outperforming in four out of the five datasets, with an improvement of up to 4.74%. A high degree of re-

		Cross-Era-Piano	Cross-Era-Orch	Cross-Era-Full	Cross-Comp-11	Cross-Comp-5
Tonal [10, 11]	**Template-based (F)**	**69.82 ±13.45**	75.25 ±12.16	70.51 ±11.57	37.41 ±19.96	50.01 ±19.77
	Tonal Complexity (G)	65.84 ±12.80	71.68 ±14.09	65.51 ±13.26	29.74 ±21.33	43.32 ±25.68
	Combined (F, G)	67.77 ±15.36	75.23 ±12.17	71.16 ±11.42	36.97 ±21.59	48.86 ±22.47
TIV	**TIV Basic (B)**	66.84 ±15.31	74.80 ±11.22	70.13 ±12.52	37.84 ±20.66	**54.75 ±17.96**
	TIV Complexity (Q)	57.99 ±17.67	69.87 ±12.87	62.18 ±13.71	29.61 ±19.77	43.16 ±19.06
	Harm. Rhythm (R)	21.26 ±28.06	28.31 ±21.51	21.65 ±26.44	7.77 ±17.64	16.47 ±19.35
	Basic + Comp. (B, Q)	65.59 ±16.77	76.68 ±10.35	71.50 ±12.03	37.82 ±20.30	53.40 ±16.85
	Combined (B, Q, R)	65.47 ±16.63	**77.19 ±9.89**	71.63 ±12.03	37.83 ±20.01	53.75 ±16.65
Combined	**F, G, B, Q, R**	64.39 ±16.27	76.70 ±11.41	73.78 ±10.80	**38.25 ±21.01**	49.72 ±19.25
Combined, no R	**F, G, B, Q**	64.78 ±15.80	76.56 ±11.29	**74.04 ±10.7**	37.89 ±21.57	50.44 ±18.72
Tonal+Transitions [13]		73.2	83.2	78.2	–	–

Table 5. Classification results for several feature groups across the *Cross-Era* and *Cross-Composer* datasets. The values on the table represent the mean classification accuracy and inter-class deviation, both expressed as percentages. For comparison, the state-of-the-art results for the *Cross-Era* dataset reported in [13] were also included.

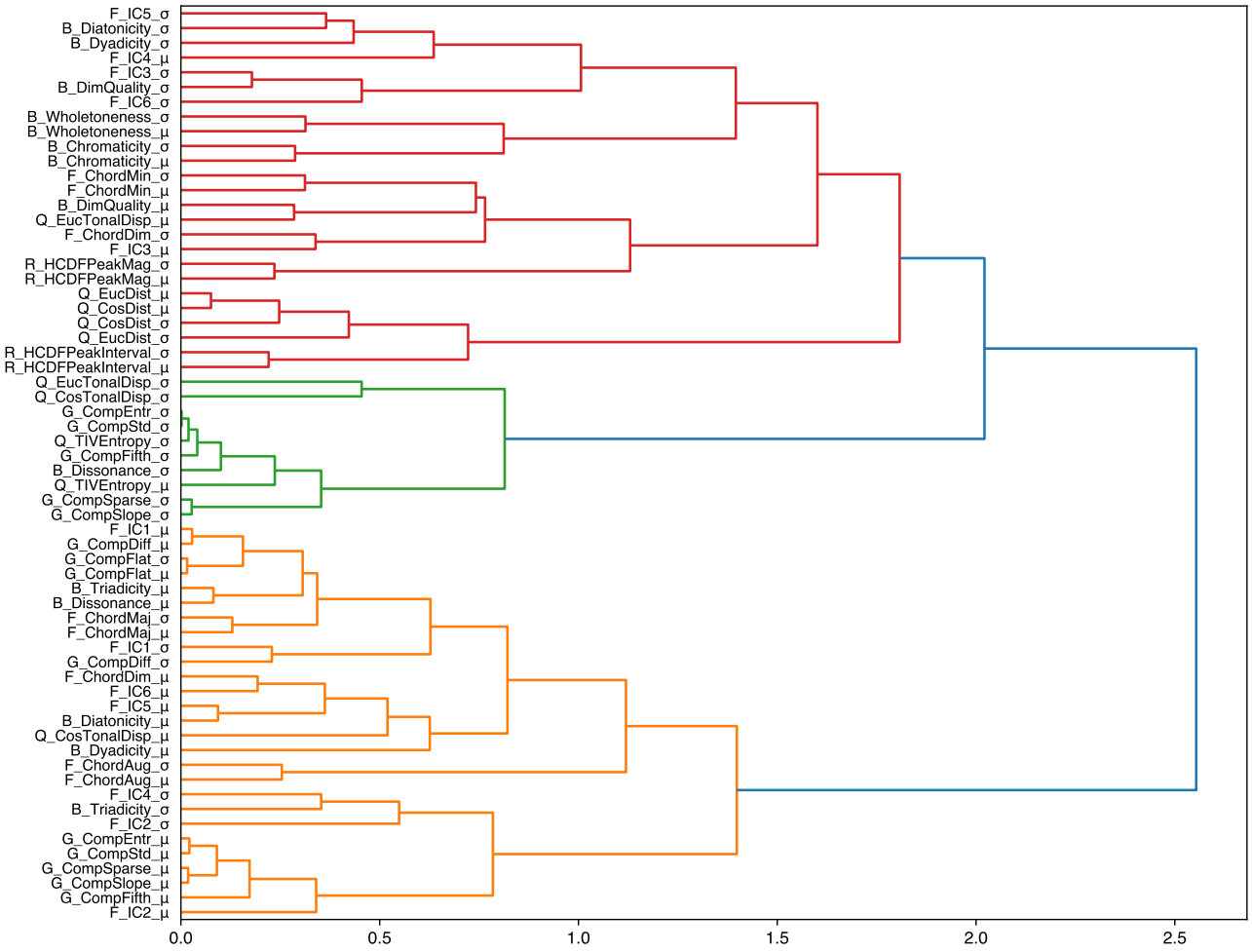

Figure 1. Hierarchical clustering of harmonic features computed at a 100ms resolution. The clustering is based on Spearman correlation distances using Ward's linkage method.

dundancy with existing state-of-the-art features has been found. However, the results suggest that new information is captured in the proposed TIV harmonic features, namely in what concerns the horizontal or long-term harmonic structure, e.g., tonal dispersion and distance between audio segments, and harmonic rhythm. While the context-sensitive segmentation strategy introduced slight improvements to classification accuracy, these do not seem to outweigh its higher computational cost.

Finally, for research reproducibility, we made the implementation of the proposed system publicly available online at `github.com/fcfalmeida/style-ident`. In future work, we may consider optimizing the segmentation strategy, expanding it to account for multiple time scales, conducting classification experiments on other musical genres, and experimenting with different machine learning approaches such as deep learning.

Acknowledgements: The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and

Fraunhofer Institut für Integrierte Schaltungen IIS. "Experimentation in music in Portuguese culture: History, contexts and practices in the 20th and 21st centuries" (POCI-01-0145-FEDER-031380) co-funded by the European Union through the Operational Program Competitiveness and Internationalization, in its ERDF component, and by national funds, through the Portuguese Foundation for Science and Technology.

6. REFERENCES

[1] K. Negus, "From creator to data: the post-record music industry and the digital conglomerates," *Media, Culture and Society*, vol. 41, no. 3, pp. 367–384, 2019.

[2] 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.

[3] S. Argamon, K. Burns, and S. Dubnov, *The structure of style: Algorithmic approaches to understanding manner and meaning*. Springer Berlin Heidelberg, 2010.

[4] J. Nam, K. Choi, J. Lee, S. Chou, and Y. Yang, "Deep learning for audio-based music classification and tagging: Teaching computers to distinguish rock from bach," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 41–51, 2019.

[5] C. Weihs, U. Ligges, F. Mörchen, and D. Müllensiefen, "Classification in music research," *Advances in Data Analysis and Classification*, vol. 1, no. 3, pp. 255–291, 2007.

[6] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, 2003, pp. 282–289.

[7] V. Tsatsishvili, "Automatic subgenre classification of heavy metal music," Master's Thesis, University of Jyväskylä, Jyväskylä, Finland, 2011.

[8] D. Gärtner, C. Zipperle, and C. Dittmar, "Classification of electronic club-music," in *Proceedings of the Deutsche Jahrestagung für Akustik (DAGA)*, Berlin, Germany, 2010.

[9] B. van 't Spijker, "Classifying classical piano music into time period using machine learning," in *Proceedings of the Twente Student Conference on IT*, Enschede, The Netherlands, 2020.

[10] C. Weiß, M. Mauch, and S. Dixon, "Timbre-invariant audio features for style analysis of classical music," in *Proceedings of the Joint Conference 40th International Computer Music Conference (ICMC) and 11th Sound and Music Computing Conference (SMC)*, Athens, Greece, 2014, pp. 1461–1468.

[11] C. Weiß and M. Müller, "Tonal complexity features for style classification of classical music," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 688–692.

[12] C. Weiß, "Computational methods for tonality-based style analysis of classical music audio recordings," Ph.D. dissertation, Ilmenau University of Technology, Ilmenau, Germany, 2017.

[13] C. Weiß, F. Brand, and M. Müller, "Mid-level chord transition features for musical style analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 341–345.

[14] G. Bernardes, D. Cocharro, M. Caetano, C. Guedes, and M. E. Davies, "A multi-level tonal interval space for modelling pitch relatedness and musical consonance," *Journal of New Music Research*, vol. 45, no. 4, pp. 281–294, oct 2016.

[15] E. Gómez, "Tonal description of music audio signals," PhD Thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

[16] M. Stein, B. Schubert, M. Gruhne, G. Gatzsche, and M. Mehnert, "Evaluation and comparison of audio chroma feature extraction methods," vol. 1, pp. 324–332, 01 2009.

[17] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 135–140.

[18] Y. Wu and W. Li, "Automatic audio chord recognition with MIDI-trained deep feature and BLSTM-CRF sequence decoding model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 355–366, 2019.

[19] A. Honingh and R. Bod, "Pitch class set categories as analysis tools for degrees of tonality," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 459–464.

[20] ——, "Clustering and classification of music by interval categories," in *Proceedings of the International Conference on Mathematics and Computation in Music*. Berlin and Heidelberg, Germany: Springer, 2011, pp. 346–349.

[21] F. Lerdahl, "Tonal Pitch Space," *Music Perception*, vol. 5, no. 3, pp. 315–349, 1988.

[22] G. Gatzsche, M. Mehnert, D. Gatzsche, and K. Brandenburg, "A symmetry based approach for musical tonality analysis," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Vienna, Austria, 2007, pp. 207–210.

[23] E. Chew, "Towards a mathematical model of tonality," PhD Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2000.

[24] R. L. Cohn, "Neo-riemannian operations, parsimonious trichords and their tonnetz representations," *Journal of Music Theory*, vol. 41, no. 1, pp. 1–66, 1997.

[25] C. Viaccoz, D. Harasim, F. C. Moss, and M. Rohrmeier, "Wavescapes: A visual hierarchical analysis of tonality using the discrete fourier transform," *Musicae Scientiae*, 2022.

[26] G. Bernardes, M. E. Davies, and C. Guedes, "A hierarchical harmonic mixing method," in *International Symposium on Computer Music Multidisciplinary Research*. Springer, 2017, pp. 151–170.

[27] A. Ramires, G. Bernardes, M. Davies, and X. Serra, "TIV.lib: an open-source library for the tonal description of musical audio," *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-20), Vienna, Austria*, no. September, pp. 304–309, 2020.

[28] E. Wold and J. Tenney, "A history of consonance and dissonance," *Computer Music Journal*, vol. 13, p. 94, 1989.

[29] J. L. Snyder, "Entropy as a measure of musical style: The influence of a priori assumptions," *Music Theory Spectrum*, vol. 12, no. 1, pp. 121–160, 1990.

[30] E. Amiot, "Entropy of fourier coefficients of periodic musical objects," *Journal of Mathematics and Music*, vol. 15, no. 3, pp. 235–246, 2021.

[31] G. Haydon and M. F. Bukofzer, "Music in the Baroque Era from Monteverdi to Bach," *The Journal of Aesthetics and Art Criticism*, vol. 7, no. 3, p. 262, mar 1949.

[32] P. Ramoneda and G. Bernardes, "Revisiting harmonic change detection," in *149th Audio Engineering Society Convention 2020, AES 2020*, oct 2020. [Online]. Available: https://www.researchgate.net/publication/342563614

[33] D. P. Ellis, C. V. Cotton, and M. I. Mandel, "Cross-correlation of beat-synchronous representations for music similarity," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2008, pp. 57–60.

[34] J. P. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signals," in *ISMIR 2005 - 6th International Conference on Music Information Retrieval*, 2005, pp. 304–311.

[35] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, London, UK, 2005, pp. 628–633.

DISTORTION AUDIO EFFECTS: LEARNING HOW TO RECOVER THE CLEAN SIGNAL

Johannes Imort[♮] **Giorgio Fabbro**[♯] **Marco A. Martínez-Ramírez**[♭]
Stefan Uhlich[♯] **Yuichiro Koyama**[♭] **Yuki Mitsufuji**[♭]
[♮]RWTH Aachen University, Germany [♯]Sony Europe B.V., Stuttgart, Germany
[♭]Sony Group Corporation, Tokyo, Japan
johannes.imort@rwth-aachen.de, giorgio.fabbro@sony.com

ABSTRACT

Given the recent advances in music source separation and automatic mixing, removing audio effects in music tracks is a meaningful step toward developing an automated remixing system. This paper focuses on removing distortion audio effects applied to guitar tracks in music production. We explore whether effect removal can be solved by neural networks designed for source separation and audio effect modeling.

Our approach proves particularly effective for effects that mix the processed and clean signals. The models achieve better quality and significantly faster inference compared to state-of-the-art solutions based on sparse optimization. We demonstrate that the models are suitable not only for declipping but also for other types of distortion effects. By discussing the results, we stress the usefulness of multiple evaluation metrics to assess different aspects of reconstruction in distortion effect removal.

1. INTRODUCTION

With the emergence of musical recordings, audio effects have become indispensable in the music production process. They are used by musicians as a creative tool to alter the sound of their instruments, and by sound engineers to craft a balanced mix from multiple recording tracks [1].

For the task of mixing and automatic remixing [2], the *dry* (i.e., unprocessed) source tracks are required. Given the recent advances in automatic mixing [3, 4] and music source separation (MSS) [5], a system could facilitate the adjustment of a stereo mixture to the taste and preferences of the user similar to [6]. However, when separating sources with a system trained on music stems (e.g., `MUSDB18` [7]) the mixing process is not considered, and, hence, the output of such a system contains the *wet* (i.e., processed) signal. As nonlinear distortion is one of the most commonly used effects for electric instruments, this work focuses on musical distortion effects that are used for aesthetic means (e.g., guitar overdrive/distortion pedals) and applied in the process of mixing (e.g., tape saturation). These distortion effects result in added harmonics, intermodulation distortion, and a compressed sound [8].

This paper investigates different deep neural network (DNN) approaches regarding their applicability to the audio effect removal problem. Our contributions can be summarized as follows:

- We show that recovering the clean signal from clipped or overdriven guitar signals can be efficiently solved with neural networks designed for source separation. The models achieve high quality and fast inference in contrast to solutions based on sparse optimization.
- We found that the superior performance of the models evaluated on the de-overdrive task can be traced back to superimposing the overdriven signal with the clean signal. We show that the architectures are suitable not only for declipping but also for other types of distortion effects.
- By discussing the results, we highlight that the metrics under evaluation prove beneficial in measuring different aspects of the reconstruction and can be advised for further investigations.

This work is organized as follows: Sec. 2 gives a formal introduction to audio effect removal and introduces the types of distortions that were used throughout this study. Sec. 3 briefly discusses previous work on iterative and DNN-based declipping approaches. The methods under evaluation are outlined in Sec. 4. Sec. 5 describes the data that were used for training, reports details about the experimental setup, and presents the chosen objective evaluation metrics. In Sec. 6, we evaluate the results of the comparative study of four different neural network architectures on the task of distortion removal in guitar signals for the SoX overdrive implementation. Then, we compare the same architectures on the declipping task to one state-of-the-art declipping algorithm using guitar signals as well as generic music signals. Lastly, Sec. 7 gives a conclusion and presents an outlook for future work. Audio examples are available online at *joimort.github.io/distortionremoval/*.

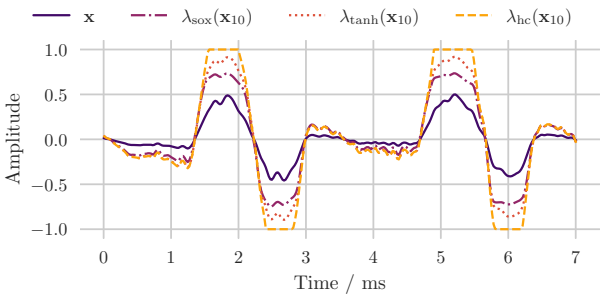

Figure 1. Example of different distortion types applied to a guitar signal. The input signal $x(k)$ is amplified before being modified by a wave-shaper function.

2. PROBLEM FORMULATION

We introduce audio effect removal as the task of recovering the original discrete audio signal $\mathbf{x} \in \mathbb{R}^n$ from the processed discrete signal $\mathbf{y} \in \mathbb{R}^n$, which is obtained by applying the possibly nonlinear and time-varying function f to the signal $\mathbf{x}$:

$$\mathbf{y} = g(\mathbf{x}) = \alpha f(\mathbf{x}) + (1-\alpha)\mathbf{x}, \tag{1}$$

with $\alpha \in [0, 1]$ denoting the weight of the wet signal, and g the summation function of the dry and wet signal. The goal is to find an estimation of the original signal $\hat{\mathbf{x}}$ by estimating the inverse function $\hat{\mathbf{x}} = \hat{g}^{-1}(\mathbf{y})$.

Generally, distortion effects clip the input signal and can be divided into systems that apply hard-clipping or soft-clipping. As this work focuses on both types of distortion, we introduce the following generic formulation of a wave-shaper that maps the amplified signal $\mathbf{x}_\gamma = 10^{\frac{\gamma}{20}}\mathbf{x}$ to a fixed range:

$$f(\mathbf{x}) = \lambda(\mathbf{x}_\gamma) \quad \text{with} \quad \lambda : \mathbb{R} \to [-\theta_\text{c}, \theta_\text{c}]. \tag{2}$$

Here, γ denotes the gain in decibels, λ the arbitrary wave-shaper function, and θ_c the fixed clipping threshold. For the case of hard-clipping, λ is defined as:

$$\lambda_{\text{hc}}(x_{\gamma,k}) = \begin{cases} x_{\gamma,k}, & \text{if } |x_{\gamma,k}| \leq \theta_\text{c} \\ \theta_\text{c}\text{sgn}(x_{\gamma,k}) & \text{otherwise,} \end{cases} \tag{3}$$

with sgn denoting the sign function, and $x_{\gamma,k}$ the k-th time sample of the amplified signal. Fig. 1 highlights the difference between different distortion types. Hard-clipping cuts off the amplitude when it exceeds a defined threshold (as typical for saturation in digital signal processing). Soft-clipping (e.g., $\lambda_{\tanh}(x_{\gamma,k}) = \tanh(x_{\gamma,k})$) gradually applies a smooth transition before reaching a fully saturated state (as typical for saturation in analog amplifiers).

Modeling the characteristics of distortion pedals in reality is more complex: [9] provides an overview on different methods and discusses DNN-based approaches. In order to simplify the problem for our investigation, we focus on wave-shaping. The overdrive algorithm of the audio editing software SoX [10] serves as an example of soft-clipping (λ_{sox}) but, in contrast to (2), it is dependent on previous samples. Furthermore, it blends the wet signal with the dry one ($\alpha < 1$).

3. RELATED WORK

To the best of our knowledge, there has been no previous research on distortion audio effect removal. Therefore, this section outlines the most relevant iterative and DNN-based declipping approaches since declipping is a special case of distortion audio effect removal.

3.1 Iterative Declipping Methods

Previous research on approaches to declipping has focused mainly on unsupervised algorithms that recover the signal under the assumption of a generic regularization such as signal sparsity [11]. Usually, these approaches target the hard-clipping case only (see (3)). While early approaches were based on auto-regressive models, (e.g., [12]) recent state-of-the-art methods evolved by combining ideas from inverse problems and sparse regularization [13].

Recently, [11] and [13] discussed popular declipping algorithms. One of the current state-of-the-art methods is A-SPADE, which will serve as a baseline for this study. The algorithm is briefly introduced in Sec. 4.

3.2 DNN-Based Declipping

In contrast to iterative algorithms, to date, there are only few contributions comprising supervised DNNs.

Kashani *et al.* [14] introduced a declipping method based on the U-Net architecture [15]. It operates on magnitude spectrograms while the waveform of the output is obtained by reusing the phase information from the distorted input signal. The system is trained and evaluated on pairs of hard-clipped and clean speech samples.

Mack and Habets [16] proposed an architecture comprising a BLSTM-based deep filtering method that works on complex spectrograms and hence also considers phase information. Similar to [14], they train the system on speech data only. Unlike any other approach, they not only investigate the system on the hard-clipping case but also on the soft-clipping case.

Tanaka *et al.* recently proposed APPLADE [17], a declipping method that takes advantage of the sparse optimization techniques described above together with deep learning. Accordingly, they embed a DNN in the iterative algorithm to enhance the thresholding operation. They report a slightly higher performance than previous algorithms, better robustness to mismatches between training and test data, and faster inference.

4. METHODS

This section describes four neural network architectures that we selected from the literature and evaluated on the distortion removal task. We approach the distortion effect removal problem from the perspective of filtering the added harmonics and intermodulation distortion, similar to [16]. Therefore, we include one model from the domain