Robustness of Features and Classification Models on Degraded Data Sets in Music Classification

Igor Vatolkin

Abstract There exists a large number of supervised music classification tasks: Recognition of music genres and emotions, playing instruments, harmonic and melodic properties, temporal and rhythmic characteristics, etc. In recent years, many studies were published in that field, which are either focused on complex feature engineering or application and tuning of classification algorithms. However, less work is done on the evaluation of model robustness, and music data sets are often limited to music with some common characteristics, so that the question about the generalisation ability of proposed models usually remains unanswered. In this study, we examine and compare the classification performance of audio features and classification models when applied for recognition of genres and instruments on music data sets which were degraded by means of techniques available in the Audio Degradation Toolbox including attenuation, compression, live and vinyl recording degradations, and addition of noise.

Igor Vatolkin
Department of Computer Science
TU Dortmund University, August-Schmidt-Strasse 4, 44227 Dortmund
✉ igor.vatolkin@tu-dortmund.de

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 5, No. 1, 2018

DOI: 10.5445/KSP/1000087327/22
ISSN 2363-9881

1 Introduction and Related Work

The goal of supervised music classification is to assign music pieces or their parts to semantic categories like genres, styles, personal preferences, emotions, playing instruments, chords, etc. As for any other supervised classification task, the choice of the training set is crucial: Data sets which are small and limited, e.g., to a few genres or a particular recording practice, lead to less robust models with poor generalisation performance. The effect of limited training sets was addressed in some studies on image recognition. For instance, the addition of perturbations (Papernot et al, 2017) or the change of brightness (Hosseini et al, 2017) lead to a significant decrease of classification performance.

The process of data augmentation generally modifies training data retaining the same annotations. Examples of data augmentation methods in computer vision are inversion, rotation, or stretching. For music signals, one may apply, for instance, transposition, change of loudness, or addition of noise, see also Mauch and Ewert (2013) and McFee et al (2015). The first benefit is that training sets become more robust. Another advantage is that the amount of training data can be largely extended without further manual efforts for labelling.

Not so much work has been done to measure the impact of augmented sets in music. Beyond the presentation of the Audio Degradation Toolbox, Mauch and Ewert (2013) showed how degradations can affect the performance of several applications including audio identification, score-to-audio alignment, beat-tracking, and chord detection. The Audio Degradation Toolbox was also applied for the measurement of degradations' impact on the classification of music genres using several classification methods in Clar (2014), for the measurement of robustness of several feature groups for the estimation of music similarity in Panteli and Dixon (2016), and for the recognition of artists in Eghbal-zadeh and Widmer (2016). The performance of the system on latin music genre recognition significantly dropped when the test set was moderately time-stretched by 32 scales in [0.85,1.15] (Sturm, 2016). Schlüter and Lehner (2018) report a significant decrease of performance of vocal recognition with Convolutional Neural Networks (CNN) for pieces with a changed loudness. In the earlier work Schlüter and Grill (2015), the impact of CNN performance was examined with regard to different augmentation methods including filters and the change of pitch and tempo, with a conclusion that not all augmentation

methods were helpful. The transposition was applied to chord prediction in Korzeniowski et al (2018) showing an increase of test performance with an augmented training data set.

In summary, many studies which measure the effect of data augmentation are either limited to individual classification methods like neural networks or individual tasks like genre recognition–and in most cases do not explicitly compare several feature sets and classifiers with regard to their sensitivity to data augmentation for different music classification tasks by the means of statistical tests.

In this work, we measure the reduction of classification performance after the application of 12-14 degradation methods for 4 classification methods and 3-4 feature sets applied to 6 genre recognition and 20 instrument recognition tasks. Section 2 lists data augmentation methods used in our experiments. Section 3 describes the experiments. The results are discussed in Section 4, including statistical comparison of degradation methods to the baseline and the evaluation of differences in performance reduction for individual classifiers and feature sets. Section 5 provides conclusions and ideas for future work.

2 Degradations

The following audio degradation methods available within the Audio Degradation Toolbox Mauch and Ewert (2013) are used for data augmentation and test of the robustness of classification models: the reduction of loudness to 20 %, 40 %, 60 %, and 80 %, live recording, strong compression, vinyl recording, four different noises with 20 dB signal-to-noise ratio (white, pink, blue, violet), radio broadcast, smartphone recording, and “old dusty” recording.

Although the changes in sound can easily be perceived by human listeners after the degradation, the spectrum is not changed to a very strong extent. Figure 1 illustrates the impact of three degradations applied on a chord for frequencies below 1500 Hz.

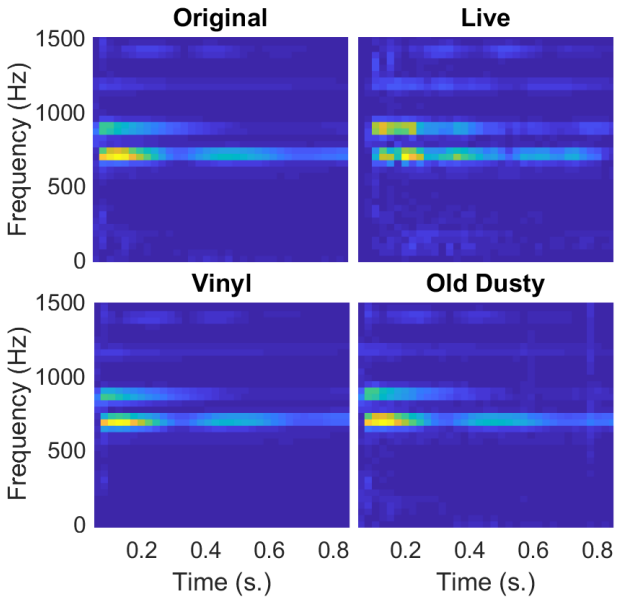

Figure 1: Impact of three degradations on the spectrum (0 - 1500 Hz) of the artificially mixed chord A5 (acoustic guitar) - F5 (bouzuki) - A5 (piano).

3 Experimental Setup

In the following, we describe the setup of experiments on measuring the robustness of classification models for classification of genres and instruments.

3.1 Classification Tasks and Data Sets

The first group of binary classification tasks is to recognise genres for classical and popular music pieces. The corresponding data set is based on 240 music pieces from sets OS120 and TAS120 of our internal database[1] with 90 Pop/Rock, 30 Classical, 30 Electronic, 30 Jazz, 30 Rap, and 30 R'n'B tracks. For the training of classification models, balanced data sets of 30 tracks from TAS120

[1] https://ls11-www.cs.tu-dortmund.de/rudolph/mi#music_test_database, accessed 31.01.2020.

were selected: The first half with "positive" tracks (belonging to the genre to recognise), and the second half with "negative" tracks equally distributed across the remaining five genres (3 tracks per genre). As in our previous studies like Vatolkin et al (2016), we use small training sets with respect to a real-world situation of a listener who prefers to define a personal category or genre with a limited number of tracks, and has less manual annotation efforts. After the training of models with different features and classifiers (see Section 3.2), we measure the classification performance on the validation set compiled with 120 tracks from the OS120 set. Then, the models are validated on 14 degraded OS120 variations.

The second group of binary classification tasks is to classify instruments in polyphonic mixtures of three or four instrument samples. Here, we use two data sets from Vatolkin and Rudolph (2018) which together contain 6000 mixtures. Each mixture contains at least one Western instrument sample taken from MUMS (Eerola and Ferrer, 2008), RWC (Goto et al, 2003), and the University of Iowa[2] databases and at least one ethnic instrument sample from Ethno World 5 Professional & Voices[3]. 3000 tracks are used to train classification models, 3000 for validation. Again, we apply degradations on the validation set. However, we use only 12 of the 14 mentioned in Section 2, because radio broadcast and smartphone recording could not be applied for short audio mixtures of several seconds. The Western instruments to recognise are acoustic guitar, cello, electric guitar, flute, piano, trumpet, viola, and violin. The ethnic instruments to recognise are balalaika, bandura, banjo framus, banjolin, bawu, dilruba, dung dkar, egyptian fiddle, erhu, fujara, melodica, and scale changer harmonium.

3.2 Features and Classification Methods

Table 1 provides an overview of audio features used in the study. All features were extracted with open source Java framework AMUSE (Vatolkin et al, 2010) with further integrated toolboxes (Chroma Toolbox (Müller and Ewert, 2011), jAudio (McKay, 2010), MIR Toolbox (Lartillot and Toiviainen, 2007), NNLS Chroma (Mauch and Dixon, 2010), and Yale (Mierswa et al, 2006)). For the

[2] http://theremin.music.uiowa.edu/MIS.html, accessed 31.01.2020.

[3] http://www.bestservice.de, accessed 31.01.2020.

genre classification task, feature values are stored only from the frames between onset events, which are assumed to have a more stable sound. The onset events were previously estimated with MIR Toolbox (Lartillot and Toiviainen, 2007). Then, the mean and the standard deviation for larger classification frames of 4 s with 2 s overlap are estimated and processed as individual features for model training. For instrument recognition from audio mixtures (chords), feature values are stored from the middle of the attack phase, onset frame, and the middle of the release phase, respectively. For instance, 13 original MFCC (Mel Frequency Cepstral Coefficient) dimensions lead to 26 dimensions for genre recognition and 39 dimensions for instrument recognition.

Table 1: Audio feature sets. Column "Size" indicates the number of dimensions after feature processing (see the text).

	Genres		**Instruments**	
Group	**Examples**	**Size**	**Examples**	**Size**
MFCCs	MFCCs	26	MFCCs	39
Timbre	*MFCCs* + spectral characteristics (centroid, bandwidth, etc.), low energy, tristimulus, phase domain features, etc.	256	*Same as for genres* + equivalent rectangular bandwidth (ERB) characteristics	474
Harmony	Fundamental frequency, chroma, bass chroma, major/minor alignment, key strengths, etc.	304	-	-
All	*Timbre* + *harmony* + fluctuation patterns, rhythmic clarity, etc.	736	*Timbre*, chroma, spectral peak characteristics, fluctuation patterns, etc.	615

Classification models are built with four classification methods available in RapidMiner (Hofmann and Klinkenberg, 2013): naive Bayes (NB), decision tree C4.5, random forest (RF), and linear support vector machine (SVM). For genre recognition, the tracks are assigned to labels by majority voting based on predictions of models for classification frames of 4 s. Because of unbalanced validation sets, evaluation is done with regard to balanced relative error m_{BRE}:

$$m_{BRE} = \frac{1}{2}\left(\frac{FN}{TP+FN} + \frac{FP}{TN+FP}\right), \tag{1}$$

where TP denotes the number of true positives (tracks belonging to a category and predicted as belonging to it), TN true negatives (tracks not belonging to a category and predicted as not belonging to it), FP false positives (tracks not belonging to a category and predicted as belonging to it), and FN false negatives (tracks belonging to a category and predicted as not belonging to it).

This measure guarantees that wrong predictions of positive and negative observations have the same impact independently of their distribution in the test set. Music tracks are assigned to genres by majority voting based on predictions of models for W classification frames of 4 s with 2 s overlap:

$$\hat{y}(\vec{x}_1, ..., \vec{x}_W) = \left\lceil \frac{\sum_{w=1}^{W} \hat{y}_w}{W} - \frac{1}{2} \right\rceil, \tag{2}$$

where $\vec{x}_w$ denotes feature vector for the w-th classification frame, $\hat{y}_w$ the label predicted for this frame, and $\hat{y}$ the label predicted for the complete track. In case some less typical segments appear in a music piece (e.g., a short bridge with strings in a Rock song), the wrong classification of corresponding time frames would not contribute to the overall error measured for a common scenario when complete tracks and not their parts are assigned to genres.

4 Analysis of Results

Figure 2 shows the balanced relative error m_{BRE} for two example categories using all features for all combinations of classifiers and degradations. The baseline performances on the original validation set ("Orig") are additionally marked with lines. Although the reduction of loudness ("Qu1"–"Qu4") produces a lower impact on the error compared to other degradations, even then the effect may be rather strong: e.g., m_{BRE} changes from 0.0591 to 0.1824 (smaller values are better) for category Classical, RF, and loudness reduced to 20 %. For several other combinations of a classifier and a category, the performance may even slightly increase after loudness reduction, as for loudness reduction/SVM/Classical. Other degradations may lead to completely unexpected results and drastically reduce the performance leading to errors close to or even exactly equal to 0.5 which corresponds to the error of the random classifier. From the first glance at these two classification tasks, the robustness of classifiers differs: SVM has a higher increase of the error than RF for Classical, but a lower increase of the error than RF for Piano.

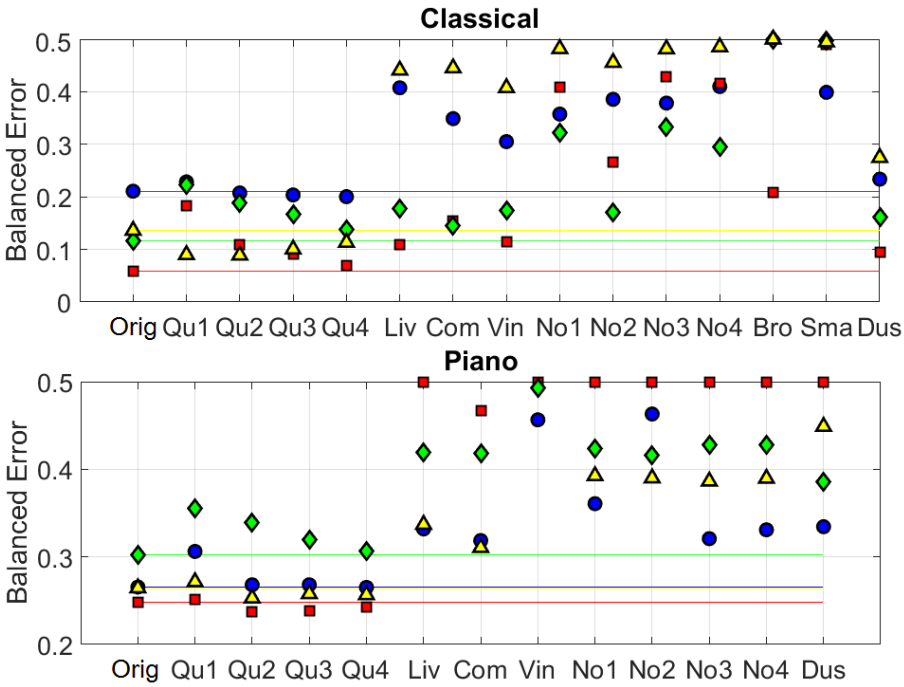

Figure 2: Balanced classification errors for two classification tasks using all features (for the genre Classical estimated after the majority voting, cf. Equ. 2). Classifiers: circles: C4.5; rectangles: RF; diamonds: NB; triangles: SVM. Orig: baseline validation error without degradations. Degradations: Qu1-Qu4: reduction of loudness to 20 %, 40 %, 60 %, 80 %; Liv: live recording; Com: strong compression; Vin: vinyl recording; No1-No4: noise (white, pink, blue, violet); Bro: radio broadcast; Sma: smartphone recording; Dus: "old dusty" recording.

For the evaluation of robustness independently of the individual category complexity, we may estimate the relative increase of m_{BRE} for degraded validation data sets in relation to m_{BRE} for the original validation set. Numbers higher than 1 would indicate the decrease of classification performance, i.e., a poor generalisation ability of the classification models. Tables 3-6 in the Appendix list this statistic averaged for 6 genre and 20 instrument categories in each case, stored separately for all combinations of a degradation, a feature set, a classifier, and a category group. As can be expected, the numbers are almost always above 1. Several cases with values below 1 correspond either to

(1) loudness reduction and compression, or

(2) vinyl recording/MFCCs/NB (genres and instruments) and smartphone recording/MFCCs/NB (genres).

For (1), we suppose that the inclusion of degradations other than loudness reduction and compression into training data may be particularly valuable to train more robust models. For (2), it is worth to mention that models built with only MFCCs usually do not achieve the best classification results, and NB has often larger errors than RF and SVM, so that the robustness in those cases does not correlate with a high classification performance.

Table 2: p-values after the application of Wilcoxon right-sided signed rank test for the following null hypothesis: The classification error achieved after the application of a given combination of a degratation, a feature set, and a classifier is *not* higher than for the original validation data set. The balanced classification error values are stored for all 26 categorisation tasks. Note that the values provided in the table are rounded to 3 digits after the fix point.

Degra-	**C4.5**			**RF**			**NB**			**SVM**		
dations	**MFCC**	**Timb**	**All**	**MFCC**	**Timb**	**All**	**MFCC**	**Timb**	**All**	**MFCC**	**Timb**	**All**
Quiet1	0.063	0.000	0.000	0.173	0.000	0.000	0.251	0.006	0.003	0.289	0.000	0.001
Quiet2	0.218	0.000	0.000	0.479	0.000	0.000	0.259	0.012	0.020	0.078	0.000	0.009
Quiet3	0.216	0.000	0.000	0.721	0.000	0.000	0.554	0.007	0.003	0.656	0.001	0.015
Quiet4	0.329	0.000	0.009	0.319	0.000	0.001	0.470	0.015	0.007	0.188	0.003	0.017
Live	0.000	0.000	0.000	0.000	0.000	0.000	0.028	0.000	0.000	0.008	0.042	0.010
Compr	0.118	0.000	0.000	0.295	0.004	0.004	0.648	0.000	0.000	0.656	0.723	0.571
Vinyl	0.001	0.000	0.000	0.001	0.000	0.000	0.936	0.000	0.000	0.016	0.324	0.081
Noise1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016	0.020	0.028
Noise2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.036	0.027
Noise3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.039	0.025	0.030
Noise4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.055	0.044	0.036
Dusty	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.047	0.002	0.001

For a more reliable statistical evaluation of the degradations' impact on music classification, we compare the mean increase of m_{BRE} to the baseline (no increase) by means of Wilcoxon right-sided signed rank test. Thus, the null hypothesis is that the performance on degraded data sets is not significantly lower (and the error higher) than the performance on the original validation data set. Table 2 contains the corresponding p-values. Please note that the numbers in Table 2 are rounded to 3 decimal places, such that values of 0.000 mean very small positive numbers above zero. Because we average for all 26 music categories (aiming to provide some general recommendations for music classification at all), 12 of 14 degradations and 3 of 4 feature sets used for both

genre and instrument recognition are taken into account. The analysis of the table leads to the following observations:

1. **Which degradations lead to a more significant decrease of performance?**

   As a simple statistic from the table, we may sum all 12 p-values (4 classifiers $\times$ 3 feature sets) for each row (degradation type): A larger number of rejected null hypotheses would lead to a smaller sum. The degradations which lead to a more significant decrease of performance for all combinations of feature sets, classifiers, and classification tasks, are "old dusty" recording (sum of the original p-values of the rows of Table 2 is equal to 0.0502), following with white noise ("Noise1", 0.0646), pink noise ("Noise2", 0.0705), and live recording (0.0886). The degradations which have the smallest impact on the decrease of performance are compression (3.0203), reduction of loudness to 60 % ("Quiet3",2.1726), vinyl recording (1.3583), and reduction of loudness to 80 % ("Quiet4", 1.3580). However, it is important to mention that these results are estimated only for genre and instrument recognition and a limited set of features and classification methods. For instance, reduction of loudness led to a significant decrease of performance for convolutional neural networks trained for recognition of vocals in Schlüter and Lehner (2018).

2. **Which classifiers are particularly sensitive to degradations?**

   Models trained with C4.5 are the most sensitive to degradations (sum of p-values in columns 2-4 of Table 2 is equal to 0.9557), followed by the random forest (1.9971), naive Bayes (3.2205), and linear support vector machine (4.0998). When we distinguish between feature groups, the order slightly varies. For MFCCs, the most sensitive classifier is again C4.5 (sum of p-values in column 2 is equal to 0.9449), followed by the RF (1.9871), SVM (2.0547), and NB (3.1466). For timbre features, the most sensitive classifier is C4.5 (sum of p-values in column 3 is equal to 0.0005), followed by the RF (0.0048), NB (0.0403), and SVM (1.2204). For all features (and hence the most powerful feature set), the order is RF (0.0052), C4.5 (0.0103), NB (0.0336), SVM (0.8247). As a conclusion, it seems to be that C4.5 and RF are more sensitive and NB and SVM are less sensitive to degradations.

3. **Which features are particularly sensitive to degradations?**

   The sensitivity to degradations increases with a growing size of the feature set: MFCCs are least sensitive to degradations (sum of p-values in columns 2, 5, 8, 11 is equal to 8.1333), followed by the timbre features (1.2660) and all features (0.8739). Although MFCCs are very often used in music classification tasks, the best classification results are usually achieved with more features, as, for example, proven in our previous experiments (Vatolkin et al, 2016), for which the MFCC set was the best only for the recognition of the music genre Rap. This means, that an extension of the feature set with more powerful and distinctive audio characteristics does not bring only advantages without side effects. For instance, the danger of reduced classification performance for degraded data sets may increase.

5 Conclusions and Future Work

In our study, we have applied various audio degradation types to the data set used for the validation of classification performance for 6 genre and 20 instrument recognition tasks. The results show that in most cases the degraded data could not be classified as well as the original data. Compression, reduction of loudness, and vinyl recording degradation led to a smaller decrease of performance (and even to an increase of performance in several cases). "Old dusty" recording, addition of noise, and live recording led to a larger decrease of performance. The increase of errors was not significantly different for genres compared to instruments in general, depending rather on a concrete classification task, degradation technique, feature set, and a classification algorithm.

Other observations are that classification models trained with linear support vector machine and naive Bayes were less sensitive to degraded data sets than models trained with decision tree C4.5 and random forest, and models trained with larger feature sets had a larger decrease of classification performance for degraded data sets.

More experiments are needed for more general and reliable recommendations, e.g., for the recognition of emotions and music styles, the classification with other methods like neural networks or ensembles, but also the validation with other performance measures like precision and recall. For a better comparison

of algorithms, we plan to extend the experiments to larger and publicly available datasets. Future work should further include degraded samples into training sets to measure a (hopeful) increase of classification performance.

Acknowledgements This work was carried out within the project 336599081 funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

References

Clar FC (2014) Impact of Audio Degradation on Music Classification. Master's thesis, Departament de Teoria del Senyal i Comunicacions, Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona. DOI: 2099.1/22605.

Eerola T, Ferrer R (2008) Instrument Library (MUMS) Revised. Music Perception 25(3):253–255. DOI: 10.1525/mp.2008.25.3.253.

Eghbal-zadeh H, Widmer G (2016) Noise Robust Music Artist Recognition Using I-Vector Features. In: Silva DF, Yeh CCM, Batista GEAPA, Keogh EJ (eds.), Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR), pp. 709–715. ISBN: 978-0-692755-06-8, URL: https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/037_Paper.pdf.

Goto M, Hashiguchi H, Nishimura T, Oka R (2003) RWC Music Database: Music Genre Database and Musical Instrument Sound Database. In: Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR), pp. 229–230.

Hofmann M, Klinkenberg R (eds.) (2013) RapidMiner: Data Mining Use Cases and Business Analytics Applications. Hofmann M, Klinkenberg R (eds.), Chapman & Hall/CRC.

Hosseini H, Xiao B, Jaiswal M, Poovendran R (2017) On the Limitation of Convolutional Neural Networks in Recognizing Negative Images. In: Chen X, Luo B, Luo F, Palade V, Wani MA (eds.), 16th IEEE International Conference on Machine Learning and Applications (ICMLA2017), Institute of Electrical and Electronics Engineers (IEEE), New York (USA), pp. 352–358. DOI: 10.1109/ICMLA.2017.0-136.

Korzeniowski F, Sears DRW, Widmer G (2018) A Large-Scale Study of Language Models for Chord Prediction. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 91–95. DOI: 10.1109/ICASSP.2018.8462285.

Lartillot O, Toiviainen P (2007) MIR in Matlab (II): A Toolbox for Musical Feature Extraction From Audio. In: Simon Dixon RT David Bainbridge (ed.), Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR), International Society for Music Information Retrieval, pp. 127–130. ISBN: 978-3-854032-18-2, URL: http://ismir2007.ismir.net/proceedings/ISMIR2007_p127_lartillot.pdf.

Mauch M, Dixon S (2010) Approximate Note Transcription for the Improved Identification of Difficult Chords. In: J. Stephen Downie RCV (ed.), Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR), International Society for Music Information Retrieval, pp. 135–140. ISBN: 978-9-039353-81-3, URL: http://ismir2010.ismir.net/proceedings/ismir2010-25.pdf.

Mauch M, Ewert S (2013) The Audio Degradation Toolbox and Its Application to Robustness Evaluation. In: Alceu de Souza Britto Jr. SD Fabien Gouyon (ed.), Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR), International Society for Music Information Retrieval, pp. 83–88. ISBN: 978-0-615900-65-0, URL: http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/145_Paper.pdf.

McFee B, Humphrey EJ, Bello JP (2015) A Software Framework for Musical Data Augmentation. In: Müller M, Wiering F (eds.), Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR), International Society for Music Information Retrieval, pp. 248–254. ISBN: 978-8-460688-53-2, URL: http://ismir2015.uma.es/articles/228_Paper.pdf.

McKay C (2010) Automatic Music Classification with jMIR. PhD thesis, Department of Music Research, Schulich School of Music, McGill University.

Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T (2006) YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Eliassi-Rad T, Ungar LH, Craven M, Gunopulos D (eds.), Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Association for Computing Machinery (ACM), New York (USA), pp. 935–940. DOI: 10.1145/1150402.1150531.

Müller M, Ewert S (2011) Chroma Toolbox: Matlab Implementations for Extracting Variants of Chroma-Based Audio Features. In: Klapuri A, Leider C (eds.), Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), International Society for Music Information Retrieval, pp. 215–220. ISBN: 978-0-615548-65-4, URL: http://ismir2011.ismir.net/papers/PS2-8.pdf.

Panteli M, Dixon S (2016) On the Evaluation of Rhythmic and Melodic Descriptors for Music Similarity. In: Mandel MI, Devaney J, Turnbull D, Tzanetakis G (eds.), Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR), International Society for Music Information Retrieval, pp. 468–474. ISBN: 978-0-692755-06-8, URL: https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/161_Paper.pdf.

Papernot N, McDaniel PD, Goodfellow IJ, Jha S, Celik ZB, Swami A (2017) Practical Black-Box Attacks Against Machine Learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (AsiaCCS), Association for Computing Machinery (ACM), New York (USA), pp. 506–519. DOI: 10.1145/3052973.3053009.

Schlüter J, Grill T (2015) Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks. In: Müller M, Wiering F (eds.), Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR), International Society for Music Information Retrieval, pp. 121–126. ISBN: 978-8-460688-53-2, URL: http://ismir2015.uma.es/articles/264_Paper.pdf.

Schlüter J, Lehner B (2018) Zero-Mean Convolutions for Level-Invariant Singing Voice Detection. In: Gómez E, Hu X, Humphrey E, Benetos E (eds.), Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), International Society for Music Information Retrieval, pp. 321–326. ISBN: 978-2-954035-12-3, URL: http://ismir2018.ircam.fr/doc/pdfs/189_Paper.pdf.

Sturm BL (2016) The "Horse" Inside: Seeking Causes Behind the Behaviors of Music Content Analysis Systems. Computers in Entertainment 14(2):3–1–3:32. DOI: 10.1145/2967507.

Vatolkin I, Rudolph G (2018) Comparison of Audio Features for Recognition of Western and Ethnic Instruments in Polyphonic Mixtures. In: Gómez E, Hu X, Humphrey E, Benetos E (eds.), Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), International Society for Music Information Retrieval, pp. 554–560. ISBN: 978-2-954035-12-3, URL: http://ismir2018.ircam.fr/doc/pdfs/139_Paper.pdf.

Vatolkin I, Theimer W, Botteck M (2010) AMUSE (Advanced MUSic Explorer) - A Multitool Framework for Music Data Analysis. In: Downie JS, Veltkamp RC (eds.), Proceedings of the 11th International Society on Music Information Retrieval Conference (ISMIR), International Society for Music Information Retrieval, pp. 33–38. ISBN: 978-9-039353-81-3, URL: http://ismir2010.ismir.net/proceedings/ismir2010-8.pdf.

Vatolkin I, Bonnin G, Jannach D (2016) Comparing Audio Features and Playlist Statistics for Music Classification. In: Wilhelm AF, Kestler HA (eds.), Analysis of Large and Complex Data, Springer, Cham (Switzerland), pp. 437–447. DOI: 10.1007/978-3-319-25226-1_37.

Appendix

Table 3: Mean increase and standard deviation of m_{BRE} for various degradations and feature sets related to baseline m_{BRE} for the original validation set, averaged for 6 genre and 20 instrument categories, for classification models created with C4.5.

Degra-	Genres				Instruments		
dations	MFCCs	Timbre	Harmony	All	MFCCs	Timbre	All
Quiet1	1.02 ± 0.04	1.31 ± 0.44	1.19 ± 0.22	1.01 ± 0.18	1.02 ± 0.06	1.58 ± 0.68	1.57 ± 0.70
Quiet2	1.01 ± 0.03	1.16 ± 0.29	1.14 ± 0.14	0.96 ± 0.15	1.01 ± 0.03	1.47 ± 0.64	1.47 ± 0.65
Quiet3	1.03 ± 0.08	1.11 ± 0.20	1.10 ± 0.09	0.97 ± 0.12	1.00 ± 0.02	1.33 ± 0.56	1.32 ± 0.54
Quiet4	1.05 ± 0.10	1.01 ± 0.05	1.03 ± 0.04	0.96 ± 0.05	1.00 ± 0.02	1.14 ± 0.25	1.13 ± 0.22
Live	1.11 ± 0.24	1.24 ± 0.44	1.22 ± 0.15	1.28 ± 0.40	1.08 ± 0.11	1.42 ± 0.47	1.38 ± 0.30
Compr	1.02 ± 0.07	1.07 ± 0.17	1.07 ± 0.10	1.09 ± 0.28	1.00 ± 0.03	1.24 ± 0.36	1.18 ± 0.20
Vinyl	1.11 ± 0.15	1.55 ± 0.99	1.19 ± 0.26	1.12 ± 0.25	1.06 ± 0.10	1.32 ± 0.41	1.31 ± 0.36
Noise1	1.33 ± 0.59	1.75 ± 1.57	1.12 ± 0.18	1.15 ± 0.29	1.13 ± 0.17	1.56 ± 0.74	1.53 ± 0.68
Noise2	1.18 ± 0.25	1.59 ± 1.27	1.06 ± 0.12	1.17 ± 0.33	1.11 ± 0.15	1.51 ± 0.66	1.54 ± 0.70
Noise3	1.33 ± 0.60	1.73 ± 1.52	1.11 ± 0.16	1.14 ± 0.33	1.12 ± 0.14	1.54 ± 0.76	1.46 ± 0.62
Noise4	1.31 ± 0.64	1.73 ± 1.56	1.11 ± 0.16	1.18 ± 0.38	1.11 ± 0.14	1.54 ± 0.75	1.47 ± 0.62
Broadc	1.09 ± 0.12	1.17 ± 0.20	1.20 ± 0.43	1.31 ± 0.60			
Smart	1.29 ± 0.55	1.83 ± 1.50	1.34 ± 0.36	1.24 ± 0.36			
Dusty	1.24 ± 0.24	1.25 ± 0.31	1.20 ± 0.24	1.09 ± 0.14	1.18 ± 0.18	1.42 ± 0.47	1.48 ± 0.51

Table 4: Mean increase and standard deviation of m_{BRE} for various degradations and feature sets related to baseline m_{BRE} for the original validation set, averaged for 6 genre and 20 instrument categories, for classification models created with random forest.

Degra-	Genres				Instruments		
dations	MFCCs	Timbre	Harmony	All	MFCCs	Timbre	All
Quiet1	1.03 ± 0.07	1.56 ± 0.77	1.50 ± 0.73	1.54 ± 0.78	1.01 ± 0.05	1.35 ± 0.52	1.38 ± 0.60
Quiet2	1.02 ± 0.06	1.31 ± 0.29	1.33 ± 0.45	1.22 ± 0.32	1.00 ± 0.03	1.28 ± 0.38	1.31 ± 0.45
Quiet3	1.01 ± 0.05	1.17 ± 0.10	1.17 ± 0.27	1.14 ± 0.20	1.00 ± 0.01	1.19 ± 0.25	1.21 ± 0.31
Quiet4	1.01 ± 0.03	1.06 ± 0.05	1.07 ± 0.12	1.03 ± 0.08	1.00 ± 0.01	1.07 ± 0.12	1.08 ± 0.12
Live	1.30 ± 0.42	1.48 ± 0.67	1.23 ± 0.21	1.55 ± 0.74	1.08 ± 0.13	1.32 ± 0.49	1.31 ± 0.43
Compr	1.03 ± 0.07	1.13 ± 0.30	0.98 ± 0.07	1.28 ± 0.66	1.00 ± 0.02	1.11 ± 0.24	1.09 ± 0.22
Vinyl	1.21 ± 0.33	1.32 ± 0.40	1.21 ± 0.26	1.30 ± 0.34	1.06 ± 0.12	1.25 ± 0.45	1.23 ± 0.30
Noise1	1.60 ± 1.13	1.99 ± 1.95	1.06 ± 0.07	2.14 ± 2.36	1.12 ± 0.16	1.34 ± 0.50	1.33 ± 0.40
Noise2	1.31 ± 0.45	1.58 ± 1.10	1.00 ± 0.10	1.65 ± 1.40	1.11 ± 0.16	1.32 ± 0.50	1.30 ± 0.37
Noise3	1.62 ± 1.19	2.01 ± 2.05	1.08 ± 0.07	2.16 ± 2.51	1.10 ± 0.13	1.34 ± 0.51	1.31 ± 0.37
Noise4	1.61 ± 1.22	1.97 ± 1.99	1.06 ± 0.05	2.11 ± 2.42	1.09 ± 0.12	1.34 ± 0.50	1.30 ± 0.36
Broadc	1.04 ± 0.08	1.24 ± 0.38	1.09 ± 0.18	1.47 ± 1.01			
Smart	1.53 ± 0.98	2.21 ± 2.37	1.34 ± 0.83	2.42 ± 2.90			
Dusty	1.27 ± 0.29	1.37 ± 0.26	1.32 ± 0.50	1.30 ± 0.23	1.12 ± 0.17	1.30 ± 0.47	1.32 ± 0.43

Table 5: Mean increase and standard deviation of m_{BRE} for various degradations and feature sets related to baseline m_{BRE} for the original validation set, averaged for 6 genre and 20 instrument categories, for classification models created with naive Bayes.

Degra-	Genres				Instruments		
dations	**MFCCs**	**Timbre**	**Harmony**	**All**	**MFCCs**	**Timbre**	**All**
Quiet1	1.01 ± 0.03	1.19 ± 0.28	1.20 ± 0.27	1.26 ± 0.35	1.00 ± 0.03	1.05 ± 0.10	1.06 ± 0.13
Quiet2	1.01 ± 0.04	1.15 ± 0.20	1.17 ± 0.22	1.19 ± 0.24	1.00 ± 0.02	1.03 ± 0.08	1.03 ± 0.11
Quiet3	1.00 ± 0.03	1.10 ± 0.15	1.11 ± 0.17	1.13 ± 0.16	1.00 ± 0.02	1.02 ± 0.05	1.03 ± 0.06
Quiet4	1.00 ± 0.02	1.05 ± 0.06	1.06 ± 0.07	1.07 ± 0.08	1.00 ± 0.02	1.01 ± 0.02	1.01 ± 0.03
Live	1.07 ± 0.23	1.44 ± 0.49	1.19 ± 0.20	1.49 ± 0.46	1.03 ± 0.08	1.50 ± 0.38	1.61 ± 0.48
Compr	0.98 ± 0.02	1.06 ± 0.14	1.07 ± 0.14	1.05 ± 0.11	1.00 ± 0.02	1.32 ± 0.28	1.41 ± 0.34
Vinyl	0.97 ± 0.06	1.10 ± 0.14	1.19 ± 0.38	1.14 ± 0.19	0.99 ± 0.05	1.40 ± 0.32	1.49 ± 0.39
Noise1	1.12 ± 0.17	1.59 ± 1.07	1.17 ± 0.34	1.40 ± 0.68	1.07 ± 0.07	1.51 ± 0.44	1.59 ± 0.53
Noise2	1.08 ± 0.12	1.26 ± 0.33	1.07 ± 0.14	1.17 ± 0.19	1.05 ± 0.06	1.39 ± 0.36	1.48 ± 0.44
Noise3	1.11 ± 0.15	1.58 ± 1.15	1.17 ± 0.32	1.40 ± 0.72	1.07 ± 0.09	1.51 ± 0.45	1.59 ± 0.53
Noise4	1.11 ± 0.16	1.48 ± 0.94	1.14 ± 0.31	1.32 ± 0.60	1.09 ± 0.10	1.49 ± 0.42	1.57 ± 0.51
Broadc	1.08 ± 0.17	1.70 ± 1.27	1.26 ± 0.41	2.09 ± 1.24			
Smart	0.99 ± 0.07	1.84 ± 1.18	1.58 ± 0.77	1.89 ± 1.18			
Dusty	1.14 ± 0.16	1.32 ± 0.30	1.21 ± 0.25	1.28 ± 0.22	1.14 ± 0.13	1.26 ± 0.30	1.29 ± 0.38

Table 6: Mean increase and standard deviation of m_{BRE} for various degradations and feature sets related to baseline m_{BRE} for the original validation set, averaged for 6 genre and 20 instrument categories, for classification models created with linear support vector machine.

Degra-	Genres				Instruments		
dations	**MFCCs**	**Timbre**	**Harmony**	**All**	**MFCCs**	**Timbre**	**All**
Quiet1	1.02 ± 0.05	1.19 ± 0.33	1.18 ± 0.29	1.01 ± 0.25	1.02 ± 0.09	1.31 ± 0.46	1.36 ± 0.62
Quiet2	1.01 ± 0.03	1.10 ± 0.18	1.12 ± 0.19	0.92 ± 0.19	1.01 ± 0.03	1.22 ± 0.33	1.25 ± 0.42
Quiet3	0.99 ± 0.02	1.07 ± 0.13	1.08 ± 0.11	0.93 ± 0.12	1.00 ± 0.02	1.14 ± 0.21	1.17 ± 0.29
Quiet4	1.01 ± 0.03	1.01 ± 0.07	1.02 ± 0.04	0.97 ± 0.08	1.00 ± 0.01	1.07 ± 0.10	1.07 ± 0.12
Live	1.27 ± 0.38	1.47 ± 0.39	1.16 ± 0.11	1.51 ± 0.86	1.04 ± 0.16	1.07 ± 0.31	1.13 ± 0.32
Compr	0.99 ± 0.04	1.05 ± 0.07	1.02 ± 0.05	1.42 ± 0.92	1.00 ± 0.00	0.95 ± 0.23	1.00 ± 0.29
Vinyl	1.09 ± 0.10	1.66 ± 1.17	1.25 ± 0.40	1.49 ± 0.74	1.03 ± 0.14	1.02 ± 0.37	1.08 ± 0.37
Noise1	1.53 ± 1.11	1.72 ± 1.33	1.07 ± 0.12	1.46 ± 1.02	1.04 ± 0.18	1.18 ± 0.47	1.27 ± 0.48
Noise2	1.24 ± 0.40	1.27 ± 0.39	1.03 ± 0.04	1.44 ± 0.94	1.04 ± 0.20	1.12 ± 0.36	1.19 ± 0.40
Noise3	1.55 ± 1.17	1.72 ± 1.41	1.08 ± 0.12	1.47 ± 1.02	1.03 ± 0.13	1.18 ± 0.46	1.24 ± 0.46
Noise4	1.56 ± 1.25	1.68 ± 1.30	1.05 ± 0.07	1.47 ± 1.03	1.01 ± 0.05	1.16 ± 0.43	1.23 ± 0.44
Broadc	1.04 ± 0.05	1.22 ± 0.38	1.21 ± 0.46	1.56 ± 1.05			
Smart	1.45 ± 0.70	1.93 ± 1.15	1.26 ± 0.49	1.74 ± 0.96			
Dusty	1.13 ± 0.18	1.31 ± 0.38	1.21 ± 0.24	1.29 ± 0.40	1.04 ± 0.19	1.21 ± 0.40	1.30 ± 0.40