



Audio Engineering Society Convention Paper

Presented at the 128th Convention
2010 May 22–25 London, UK

The papers at this Convention have been selected on the basis of a submitted abstract and extended precis that have been peer reviewed by at least two qualified anonymous reviewers. This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Automatic Detection of Audio Effects in Guitar and Bass Recordings

Michael Stein¹, Jakob Abeßer¹, Christian Dittmar¹, and Gerald Schuller²

¹*Fraunhofer IDMT, Ilmenau, Germany*

²*Ilmenau University of Technology, Ilmenau, Germany*

Correspondence should be addressed to Michael Stein (peter.michael.stein@hotmail.com)

ABSTRACT

This paper presents a novel method to detect and distinguish ten frequently used audio effects in recordings of electric guitar and bass. It is based on spectral analysis of audio segments located in the sustain part of previously detected guitar tones. Overall, 541 spectral, cepstral and harmonic features are extracted from short time spectra of the audio segments. Support Vector Machines are used in combination with feature selection and transform techniques for automatic classification based on the extracted feature vectors. With correct classification rates up to 100% for the detection of single effects and 98% for the simultaneous distinction of ten different effects, the method has successfully proven its capability - performing on isolated sounds as well as on multitimbral, stereophonic musical recordings.

1. INTRODUCTION

Audio effects are commonly applied to modify audio signals to achieve a perceivable alteration of sound attributes such as loudness, pitch or timbre. Especially in conjunction with instruments like electric guitar and bass they are extensively used to shape and customize the instrument's sound. Audio effects are an active research topic in audio technology, but they are scarcely addressed in semantic music analysis although automatic music annotation might

benefit from the knowledge about the presence of audio effects applied to an instrument. Automatic music transcription systems will most likely fail in automatically transcribing the rhythm or pitch of a melody that has been heavily processed with delay or modulation effects - a common practice in modern music production. It is uncertain, if an instrument classifier would identify the sound of a distorted guitar to be a guitar or rather a brass ensemble, which might be more similar in terms of spectral width for

instance. Furthermore, a high-level descriptor indicating the presence of a distorted guitar is assumed to be a valuable feature for music genre identification. In this paper, we show that automatic classification with Support Vector Machines based on extracted audio features provides a suitable approach to evaluate the presence of audio effects in recorded instrument sounds.

The remainder of this paper is organized as follows: After outlining the goals and challenges of this publication and the problems we face in Sec. 2, we provide an overview of related work in the subsequent section. We present the individual processing stages of our approach in Sec. 4, introducing a novel audio feature extraction concept based on harmonic analysis of instrument notes. In Sec. 5, we explain the performed experiments and discuss the obtained results. Finally, Sec. 6 concludes this work.

2. GOALS AND CHALLENGES

Our goal is to develop and evaluate a method which enables the detection and a distinction of ten different audio effects in recordings of electric guitar and bass. It is supposed to perform both on isolated musical sounds as well as on instrument tracks that are embedded in multitimbral, stereophonic musical recordings, e.g. a guitar solo within a song. The extraction of the relevant, effect related sound information from the mixture signal clearly poses the biggest challenge, since it also contains sound information originating from the instrument and, in the case of musical pieces, accompaniment sounds, artifacts or noise.

3. RELATED WORK

Several publications are dealing with audio effects, mostly focusing on audio production related aspects. Addressed topics include the technical principles of digital audio effects and how these affect sound quality [1, 2, 3, 4, 5, 6], emulation of analog circuitry behavior [7, 8, 9], adaptive audio effects [10, 11, 12] and novel processing techniques [13, 14, 15]. Furthermore, the timbral characteristics of selected audio effects [16, 17] as well as methods to categorize audio effects [18] have been studied. In music analysis, audio effects applied to instrument sounds have rarely been investigated yet. Classification of musical instrument sounds is mainly performed on a level

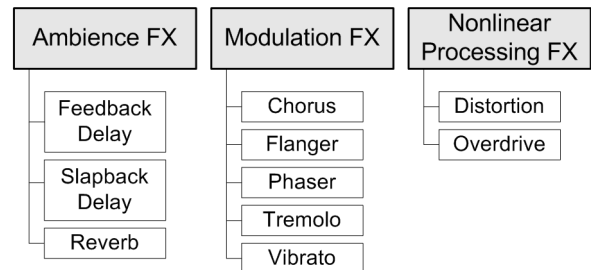


Fig. 1: Taxonomy of audio effects and how they class with effect groups.

of instrument families or instrument types. [19] provides a comprehensive review of different approaches and their results. These approaches usually neglect most of the timbral variations that may occur within the scope of a single instrument’s sound due to different playing styles, instrument design and signal processing with effects. The former of these issues have already been tackled in recent studies [20, 21, 22, 23]. To the best of our knowledge, the analysis of instrument recordings regarding the presence of audio effects has not been studied yet.

4. PROPOSED APPROACH

Although there exists a huge number of different audio effects, a few of them can be considered as de facto standards for sound processing in popular music. Regarding guitar sound shaping, the most frequently used effects can be grouped to *Ambience*, *Modulation*, *Distortion*¹ and *Filter* effects, of which the first three are treated in this paper. The audio effects that shall be detected are: *Feedback Delay*, *Slapback Delay*, *Reverb*, *Chorus*, *Flanger*, *Phaser*, *Tremolo*, *Vibrato*, *Distortion* and *Overdrive*. For detailed descriptions of these audio effects please refer to [1]. The assignment of all audio effects into the effect groups is depicted in Fig. 1.

4.1. General Structure

The method we propose is based on spectral analysis of audio segments located in the sustain part of guitar tones because they can be regarded as having a stable harmonic structure with only minor amplitude changes. The majority of audio effects has

¹In the following we will refer to this group as *Nonlinear Processing* effects to avoid confusion with the *Distortion* effect itself.

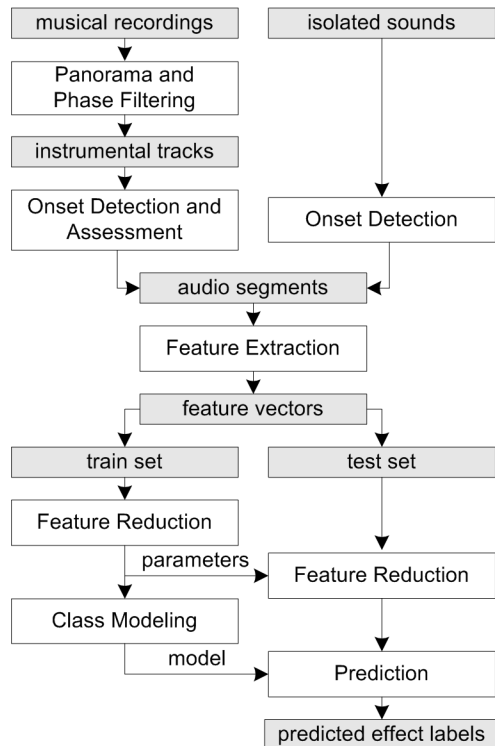


Fig. 2: Processing flow chart of the proposed method.

a time-varying behavior. By only investigating the sustain part and neglecting the attack part at the same time, we expect the extracted features to be invariant to the instrument sound to a certain extent. Thus, we assume that the detected sound variations correspond directly to the audio effect applied to the signal. In addition, we can also expect invariance to plucking styles to a certain extent, since these primarily affect the attack part of the tone [20]. Therefore, a preprocessing of the audio signal is required, which has to be adapted to the characteristics of the audio signal, i.e. we use different strategies for isolated sounds and musical recordings. The complete structure of the proposed method is shown in Fig. 2. It consists of four stages: preprocessing, feature extraction, feature reduction and classification. These will be described in detail in the following sections.

4.2. Preprocessing

For isolated sounds, which are available in single files, the preprocessing is conceivably simple. We

compute the energy envelope of the sound and obtain an initial estimate of the end of the attack part with the fixed threshold method described in [24]. The actual transition between attack and sustain part of the sound is marked by the next local maximum of the envelope. To ensure that the audio segment used for feature extraction will be fully located in the sustain part, we determine its start point behind the detected maximum.

The preprocessing for multitimbral, stereophonic musical recordings comprises three major steps: isolating the desired instrument track from the mixture, performing an onset detection and assessing the onset candidates to obtain a list of promising notes whose sustain parts can be used for analysis.

4.2.1. Panorama Filtering

To isolate the instrument track from the mixture we employ a human-assisted panorama and phase filtering method presented in [25]. It uses binary time-frequency filters to extract musical content from a stereophonic recording at a distinct panorama position. If the parameters fed into the algorithm are chosen appropriately, the resulting signal will contain the desired instrument with the rest of the mixture removed or at least well attenuated.

4.2.2. Onset Detection

To identify individual notes in the instrument track we perform an onset detection. There exist a variety of different onset detection algorithms, which basically comprise the same two steps: Derive a detection function from the signal and apply a peak-picking algorithm to it to locate the onsets [26, 27]. To derive the detection function we adapt a method based on band-wise linear prediction [28]. We perform the prediction in three sub-bands of the audio signal (0-0.5, 0.5-1, 1-2 kHz) and sum up the prediction errors of the sub-bands to obtain the detection function which is further smoothed using a 10 ms Hann window. We post-process the detection function by applying an adaptive thresholding to it. The threshold consists of a fixed and a variable value. The latter is derived from the moving average using an asymmetric rectangular window which accounts more for preceding values and therefore performs elevated thresholding subsequent to an onset.

From the detection function we obtain a set of onset candidates through conditional peak-picking, i.e.

only local maxima that meet three criteria are considered as onset candidates. First, local maxima have to have a certain height to be considered an onset candidate (*minimum strength criterion*). In this way we exclude spurious peaks. Secondly, local maxima have to have a minimal distance between them. If two maxima have a smaller distance, the stronger one is chosen as onset candidate, because we assume that not both maxima may correspond to note onsets if the distance is chosen sufficiently small, e.g. 20 ms (*minimum distance criterion*). The third criterion takes into account the requirements of the subsequent analysis. To capture the time-varying behavior of the audio effects, the sustain part to be analyzed has to have a minimal length, e.g. 500 ms. Therefore, we check every pair of consecutive onsets if it fulfills this *minimum length criterion* and if it fails, the first of the two candidates is removed from the list.

4.2.3. Onset Quality Assessment

The reduced set of onset candidates may still contain onsets whose corresponding tones are not suited to the requirements of the following feature extraction. At least three different errors can occur that have to be handled. Since only an onset and no offset detection was performed, there may be audio segments that contain short notes followed by silence. To detect these we calculate the energy envelope of the segment and check if it stays above a certain threshold all the time. Secondly, there may be multiple notes within one segment. This can happen due to valid but previously not detected onsets or weak onsets in the case of guitar specific note transition playing techniques like hammer-ons or slides. To catch previously not detected onsets, we examine the temporal envelope change for peaks which should reveal a yet not detected onset. To identify weak onsets, we use autocorrelation to compute a pitch contour and utilize its derivative to monitor if the speed of pitch change stays within predefined thresholds. Only if none of these errors occurred, the audio segment is used for further analysis. However, all the thresholds must be carefully determined to not exclude segments because of the applied audio effect. This holds especially true for delay effects which are likely to produce peaks in the temporal envelope change as well as modulation effects, which by nature will affect the pitch stability. For the ex-

periment described in this paper, we used a separate development set of instrument tracks to determine the optimal threshold values empirically.

4.3. Feature Extraction

An audio segment is first transformed into successive short time spectra. We use a sampling frequency of 44.1 kHz, frames of 8192 samples with a hopsize of 512 samples and a Hann window.

The final feature vector has a total of 541 dimensions and contains 136 spectral, 84 cepstral and 321 harmonic features, whose extraction will be described in the next sections.

4.3.1. Spectral Features

The following features are extracted frame-wise from the magnitude spectrogram: spectral centroid, spread, skewness and kurtosis, spectral flux, roll-off, slope and flatness measure. Descriptions of these and other audio features can be found in [24] and [29]. We obtain three additional feature curves through highpass filtering of spectral centroid, roll-off and slope. Furthermore, we calculate the first and second derivative of all feature curves. To characterize their value range, we calculate the mean value and standard deviation. To capture their temporal progression, we consider the feature curves as distributions and derive the statistical descriptors mean value, variance, skewness and kurtosis.

4.3.2. Cepstral Features

We apply the discrete cosine transform to the logarithmized, squared magnitude spectra to convert the spectral frames to cepstral frames and use the first ten coefficients, averaged over the whole segment, for feature extraction. These will directly contribute to the final feature vector along with their maximum value. In addition, we compute the mean value and standard deviation of the element-wise differences as well as the summed-up differences from the linear interpolated slope of the coefficients. Using the same procedure for the standard deviations of the coefficients and repeating it for the first and second derivative we obtain a total of 84 figures for the final feature vector.

4.3.3. Harmonic Features

First, we need to extract individual harmonics from the short time spectra, but we have to distinguish between monophonic and polyphonic sounds. For monophonic sounds we start with estimating the

pitch using frame-wise autocorrelation. We improve the estimation by restraining the possible range of pitches to only those that can be produced by a guitar or bass guitar in standard tuning and use the mode of all frames to get a more reliable estimation. Ideally, the harmonics would occur at the integer multiples of the fundamental frequency, but in the case of stringed instruments we have to consider the inharmonicity phenomenon, i.e. little frequency deviations caused by the stiffness of the strings [30]. Therefore, we search for a harmonic in a local frequency range around the expected position and assume the harmonic's frequency to be at the frequency bin with the highest magnitude. As we assume the guitar sound to remain stable during the course of the analyzed audio segment, we average the determined harmonic's frequencies of all frames. To avoid the challenges of multiple pitch estimation and separation of overlapping harmonics in the case of polyphonic sounds, we pursue a different strategy. Instead of trying to extract series of harmonics, we perform a peak-picking on the whitened long term average spectrum [29] to obtain a set of salient harmonics, accepting the fact that we discard useful structural information, which will have a bearing on the feature extraction.

Analysis shall reveal changes in frequency, level and shape of harmonics in their temporal progression. Hence, harmonics will not be represented by one single frequency bin, but a narrow frequency band. Within this band the actual maximum position and value as well as the band energy are computed per frame. In the following, we will refer to the resulting vectors as *harmonic feature curves*. Given the average frequency f_i of the i -th harmonic, the values of the harmonic feature curves for the m -th frame can be calculated as follows:

$$H_i^{max}(m) = \max(X_{dB}(m, \mathbf{k})) \quad (1)$$

$$H_i^{pos}(m) = \arg \max(X_{dB}(m, \mathbf{k})) \quad (2)$$

$$H_i^{en}(m) = \overline{X_{dB}(m, \mathbf{k})} \quad (3)$$

$$\mathbf{k} = f_i - \Delta k, \dots, f_i + \Delta k, \quad \Delta k < f_0/2 \quad (4)$$

where X_{dB} denotes the logarithmized, squared magnitude spectrogram, H_i^{max} the harmonic feature curve related to the maximum value of the i -th harmonic, H_i^{pos} the harmonic feature curve related to

the maximum position of the i -th harmonic, H_i^{en} the harmonic feature curve related to the band energy of the i -th harmonic and \mathbf{k} is a vector containing the frequency bin indices of the current analysis frequency band. The value for Δk has to be at least smaller than half the fundamental frequency for monophonic sounds in order not to overlap analysis bands of adjacent harmonics. For polyphonic sound it has to be even smaller due to the interleaved harmonic series that relate to different fundamental frequencies. We empirically found $\Delta k = 1/10$ to give the best tradeoff between bandwidth and number of usable analysis frequency bands.

Following this approach, several other descriptors could be extracted to further characterize aspects of harmonic sounds, skewness and kurtosis for instance, but with respect to the size of the feature vector, here we restricted analysis to just the three harmonic feature curves described above.

To obtain meaningful features from the harmonic feature curves we propose to perform spectral analysis on them. The values for frame size and hop-size have to be chosen with regard to the sampling rate of the curves which is 86 Hz in our case due to the hopsize of 512 samples at a sampling rate of 44.1 kHz for the underlying spectrogram. We used frames of 64 samples and a hopsize of 16 samples. From the resulting spectrograms, we consider only the frequency range 0...22 Hz for further analysis. This will sufficiently capture the majority of low frequency modulations and variations. The Fig. 3-6 show a few of these spectrograms for different effects applied to the same tone.

The following features are extracted frame-wise from the spectrograms: mean value, variance and standard deviation, maximum value and position, spectral centroid, spread, skewness and kurtosis as well as the steady and cumulated alternating component and their ratio. Figures for the final feature vector are derived by calculating the mean value and standard deviation to characterize the value range of the extracted features. Thus, we derive a total of 72 figures per harmonic. We performed the spectral analysis for the first ten harmonics and applied two grouping schemes to reduce the number of figures for the final feature vector. First, we simply average the figures over all harmonics. In addition, we perform a tristimulus-like grouping inspired by the

idea that the higher the order of the harmonics the more likely it is that they will be perceived rather as a group than as individual harmonics [29]. Figures of the first harmonic remain unchanged, but we average the figures of the second to fourth harmonic as well as the figures of the fifth to tenth harmonic. In case of monophonic sounds we used both grouping schemes whereas for polyphonic sounds we could only use the first. This is because we lack the information of the harmonics' order due to the applied extraction technique.

Besides the spectral analysis of the harmonic feature curves, we also analyze the temporal progression of the harmonics' levels using the curves related to the band energy H_i^{en} (see Eq. 3). For unprocessed sounds the levels will be decreasing because the guitar string is performing a damped oscillation. The use of audio effects can break this rule. This is quite obvious for delay effects which add the energy-rich attack phase to the sustained sound at a certain time instance, therefore raising the energy level. The same applies also to some modulation effects which provoke prominent level differences caused by time and frequency dependent boosts and attenuations.

We derive an ideal slope of a harmonic's levels by means of linear regression of the energy levels and compute the frame-wise differences to the real levels for the first ten harmonics. Since we use a logarithmized spectrogram, a linear slope here corresponds to exponential decay in the linear amplitude domain, which is a reasonable assumption for guitar tones. We then calculate average values and standard deviations of the difference curves and the rectified difference curves and apply the same grouping of figures as for the spectral analysis of harmonics. Furthermore, we calculate the ratio between positive and negative valued elements of the difference curves and use statistical figures to evaluate the distribution of this ratio over the harmonics.

Additionally, two harmonic features were extracted for monophonic sounds only. These are the odd-to-even harmonic ratio and the tristimulus [29].

4.4. Feature Reduction and Classification

The raw extracted features can already be used for classification but there might be correlated, redundant or irrelevant features. Hence, we insert an optional feature reduction stage in front of the classification stage. We use two algorithms which aim two

identify an optimized subset of features - namely Inertia Ratio Maximization using Feature Space Projection (IRM) [31] and Linear Discriminant Analysis (LDA) [32]. The former performs an iterative feature selection based on maximization of the ratio of between-class inertia to the total-class inertia. The latter linearly maps the feature vectors into a new, reduced feature space, guaranteeing a maximal linear separability by maximization of the ratio of between-class variance to the within-class variance. As classifier, we use renowned Support Vector Machines with a radial basis function kernel. More details on these methods can be found in [33].

5. EXPERIMENTS AND RESULTS

5.1. Data Sets

We assembled a novel data set consisting of monophonic and polyphonic guitar and bass guitar sounds processed with a variety of different audio effects. We recorded all single notes up to the 12th fret on every string of two electric guitars and bass guitars using different playing styles (finger plucking and picking with a plastic pick) and different pickup settings to cover a wide timbral range of instrument sounds. Polyphonic sounds were created by choosing appropriate fingering schemes, comprising common two note intervals as well as three and four note chords, and mixing together the associated single notes. With a digital audio workstation we processed every sample with each of the ten audio effects in at least three different settings using various effect plugins. We intend this data set, which contains a total of about 55000 samples (30 hours of recorded sounds) to be a public benchmark set for the given tasks².

To investigate the performance of the proposed method on multitimbral musical recordings we assembled a second data set using six instrumentals from different musical genres for which the single instrument tracks were available. We recorded matching guitar solos, processed them with effects and positioned all tracks in the stereo panorama prior to the stereo mix. In total, this set contains 66 accompanied guitar solos with a playtime of 45 minutes.

²See http://www.idmt.fraunhofer.de/eng/business%20areas/smt_audio_effects.htm for further information.

5.2. Five Experiments on Effect Detection

We designed five different experiments of increasing complexity, with each investigating a certain aspect of audio effect detection. Thereby, the first two experiments investigate the general feasibility while the following three explore the applicability of the proposed method.

Experiment 1 In the first experiment we check if single effects can be detected generally. The underlying intention is to figure out if the extracted features properly capture the sound characteristics of the audio effects. Therefore, we perform ten tests in a two class configuration, with the first class containing samples of the effect in question while the second class always holds the unprocessed samples.

Experiment 2 Given prior knowledge about the effect group, the objective in this experiment is to find out if the effects within this group can be distinguished from each other. Since they show functional and tonal similarities they will also share a lot of features capable to detect them, but there might be additional features which still possess enough discriminative power.

Experiments 3 and 4 In these experiments we investigate how well the effect groups or the single effects can be distinguished from each other. According to the taxonomy in Fig. 1 these experiments are performed with four and eleven classes respectively, including the unprocessed samples as an additional class.

Experiment 5 This experiment is similar to the preceding one, but now we try to gain a benefit from exploiting the hierarchy of the taxonomy. Therefore, we first determine the effect group and afterwards specify the single effect within that group. Hence, it can be regarded as a combination of the third and second experiment.

5.3. Experimental Setup

To get detailed insights into what affects the performance of the proposed method, we conducted the experiments on different subsets of the data. The first subset contains all monophonic bass guitar samples (BS-MO), the second the monophonic

guitar samples (GIT-MO). These two combined form the third subset (BS-GIT) and the fourth combines monophonic and polyphonic guitar samples (GIT-MP). The multitimbral musical recordings (MUS) make up the fifth subset.

For feature reduction, we tested all possible combinations of enabling or disabling IRM and LDA. The number of features to be selected by the IRM algorithm was varied between 20 and 160 with a step size of 20 and the number of feature dimensions after LDA transform was set to $L - 1$, where L is the number of classes.

We performed 10-fold cross-validation and an additional cross-validation where we split the data according to contextual information. For the BS-MO and GIT-MO sets this meant to combine all samples from one instrument to form either train or test set, thus exposing the influence of instrument timbre. The BS-GIT set was split into the constituting subsets. Again, exhibiting the influence of timbre it also serves as an indicator for generalization capabilities, since the guitar and bass guitar samples were processed with different effect settings. Splitting the fourth subset into monophonic and polyphonic samples obviously delivers an indicator for the influence of polyphony.

To evaluate the performance on musical recordings we use three different data sets to train the classifier while the MUS data always forms the test set. First, we use the GIT-MO set for training. We reuse it in a second run but this time we take randomly chosen snippets of music from a separate collection and add them with reduced level to the isolated sounds to account for the fact that preprocessing rather attenuates the accompanying sounds than removing them. Finally we use the MUS set for training and evaluate the outcome using a modified leave-one-out cross-validation scheme: Let the data set contain N samples. Then the test set is formed by N times choosing one sample. Instead of building the train sets from the remaining $N - 1$ samples, $N - S$ samples are chosen, where S denotes the number of samples that are located in the same musical recording the test item was taken from.

In this context, we additionally integrated a reasonable heuristic to boost the performance: For the course of a guitar solo or a musical segment in general we can assume the effect setting to remain fixed.

Data Set	Experiment 1	Experiment 2
BS-MO	98.9 (2.2)	99.3 (1.1)
GIT-MO	99.7 (0.7)	99.4 (0.9)
BS-GIT	98.9 (1.8)	98.7 (2.2)
GIT-MP	99.5 (1.4)	98.6 (2.3)

Table 1: Mean classification accuracies [%] for the first two experiments using 10-fold cross-validation, standard deviations [%] given in brackets.

Therefore, we re-evaluated the accuracy based on complete tracks by applying a majority voting strategy to the predicted effect labels.

5.4. Results and Discussion

5.4.1. Isolated Sounds

The best classification results using 10-fold cross-validation were always obtained when applying LDA while for the contextual cross-validation the best feature reduction setting differed depending on the classification task. Hence, only for the latter case the feature set leading to the highest accuracy will be indicated, implying that the indicated results for 10-fold cross-validation were always obtained applying just LDA to the data.

The results for the first and second experiment are presented in Table 1. They show that almost perfect detection and distinction could be achieved with all data sets. Regarding the first experiment, only for the effects Chorus, Phaser and Overdrive the performance degraded slightly to about 95% while all other effects achieved scores of 99% to 100%. This confirms the general feasibility of the proposed method.

The results of the third experiment, investigating the distinction of effect groups, are depicted in Table 2. In the case of 10-fold cross-validation, the results range from 94.0% for the BS-MO set to 99.0% for the GIT-MO set. For contextual cross-validation, an increased amount of false classifications could be observed for the BS-MO set. The majority of them was assigned to the class containing the unprocessed samples so this may be regarded as a problem of when to consider an effect as present related to its intensity. The same applies to the GIT-MO set but to a slighter extent. Musical signals follow a logarithmic frequency scale, so applying a linear frequency transform will provide a better frequency res-

Data Set	10 CV	Contextual CV
BS-MO	94.0	83.7 (IRM40+LDA)
GIT-MO	99.0	96.6 (LDA)
BS-GIT	95.2	82.7 (IRM120+LDA)
GIT-MP	98.2	77.6 (IRM160+LDA)

Table 2: Mean classification accuracies [%] for the third experiment, best feature set for contextual cross-validation given in brackets.

olution for sounds in higher spectral domains. This allows the features to resolve tiny differences better than in lower spectral domains and therefore keeping their discriminative power. The results of the combined BS-GIT set lie on the same level as those achieved with bass guitar samples alone. This indicates the desired invariance towards instrument timbre because otherwise much lower scores would be achieved. Regarding the influence of polyphony, the results achieved when considering it in the train set, as it is the case for 10-fold cross-validation, do not significantly differ from those obtained with monophonic guitar samples alone. In contrast, accuracy decreases notably when trying to predict audio effects in polyphonic sounds while training the classifier with monophonic ones and vice versa.

The results of the fourth experiment (Table 3), investigating the distinction of single effects, are quite comparable to those of the preceding one. Notable differences occur for the combined data sets BS-GIT and GIT-MP, especially for the contextual cross-validation. The achieved classification accuracy of the former set of 76% is the result of a bimodal distribution: All Ambience effects, Tremolo, Vibrato and Distortion effect achieve scores greater than 90% indicating that their detection and distinction is truly independent of instrument timbre. The second cluster is made up of the effects Chorus, Flanger, Phaser and Overdrive, which achieve an average accuracy of 51%. However, their false classification scatter is not evenly spread over all classes but stays within their class limits with again some samples being misclassified as unprocessed ones. This points to the low intensity issue again and also reveals a lack of discriminative power of the feature set to properly capture the sound characteristics of the involved audio effects in this scenario. The grown difference of the achieved accuracies for the GIT-MP set reinforces the

Data Set	10 CV	Contextual CV
BS-MO	93.0	84.5 (IRM120+LDA)
GIT-MO	97.7	95.7 (LDA)
BS-GIT	93.1	76.0 (LDA)
GIT-MP	95.5	63.3 (All Features)

Table 3: Mean classification accuracies [%] for the fourth experiment, best feature set for contextual cross-validation given in brackets.

need to consider monophonic and polyphonic sounds while training a classifier if one can expect both of them to be present in the data to be predicted.

Comparing the results of the third and fourth experiment one can not expect a significant increase in classification accuracy from exploiting the hierarchy of the effect taxonomy, since a false classification of the effect group can not be revised in the second classification step. In fact, the results obtained with 10-fold cross-validation showed only minor gains in accuracy between 0.2% and 1.4%.

5.4.2. Musical Recordings

In the MUS set one effect - Slapback Delay - had to be excluded from evaluation, because with the chosen segment size of 500 ms, no sufficient variety of effect settings could be maintained in the associated train data of isolated sounds, hence only 60 tracks were used. Furthermore, we only conducted the third and fourth experiment on this data because we wanted to explore the applicability of the proposed method for real world scenarios.

The results of the third experiment on this data are shown in Table 4. The best results were obtained by training the classifier with the MUS data. Applying both, feature selection and transform, 77% of the segments and even 92% of the recordings were classified correctly. This underlines the benefit of the introduced heuristic. Using the isolated sounds for training didn't gain reasonable results although adding music to the isolated sounds (GIT-MO 2) improved the accuracy. In both cases the classifier tended to assign the majority of samples to the class of modulation effects. The reason for that are artifacts introduced in the panorama and phase filtering, whose setup remains a tradeoff between preservation of signal quality of the desired track and proper attenuation of the unwanted signal. Therefore, it is

Train Set	Feature Set	Accuracy	
		Segments	Tracks
GIT-MO 1	LDA	33.7	37.5
GIT-MO 2	IRM120+LDA	38.0	47.5
MUS	IRM140+LDA	77.1	91.7

Table 4: Classification accuracies [%] for the third experiment, performed on musical data.

assumed that, if one is able to identify the critical processing and incorporate its impact in the train data or applies a less critical preprocessing the results will catch up to those obtained with the MUS data.

Accordingly, the fourth experiment was conducted only for the case of training the classifier with the MUS data. 52% of the segments and 80% of the recordings were classified correctly, again emphasizing the benefit of the majority voting strategy.

6. CONCLUSIONS

In this paper we introduced a novel method to detect and distinguish audio effects in recordings of electric guitar and bass. We showed how commonly used audio features can be adapted to capture the characteristic sound alterations caused by effect processing and presented a new approach to analyze and characterize aspects of harmonic sounds. Furthermore, we suggested a preprocessing strategy that enabled the method not only to perform on isolated instrument sounds but also on multitimbral, stereophonic musical recordings. The obtained results indicate that the proposed approach might be a valuable enhancement for existing music analysis systems. Future steps will focus on extending the method to also detect multiple, cascaded effects.

7. ACKNOWLEDGEMENTS

This work has been partly supported by the German research project GlobalMusic2One funded by the Federal Ministry of Education and Research (BMBF-FKZ: 01/S08039B). Additionally, the Thuringian Ministry of Economy, Employment and Technology supported this research by granting funds of the European Fund for Regional Development to the project Songs2See, enabling transnational cooperation between Thuringian companies and their partners from other European regions.

8. REFERENCES

- [1] Udo Zölzer, editor. *DAFX - Digital Audio Effects*. John Wiley & Sons, Chichester, 2002.
- [2] Esa Piirilä, Tapio Lokki, and Vesa Välimäki. Digital signal processing techniques for non-exponentially decaying reverberation. In *Proc. of the 1st COST-G6 Workshop on Digital Audio Effects (DAFx)*, 1998.
- [3] David Oboril, Miroslav Balik, Jiri Schimmel, Zdenek Smekal, and Petr Krkavec. Modelling digital musical effects for signal processors, based on real effect manifestation analysis. In *Proc. of the COST G-6 Conference on Digital Audio Effects (DAFx)*, 2000.
- [4] Jon Dattorro. Effect design: Part 1 reverberator and other filters. *Journal of the Audio Engineering Society*, 45(9):660–684, 1997.
- [5] Jon Dattorro. Effect design. part 2: Delay-line modulation and chorus. *Journal of the Audio Engineering Society*, 45(10):764–788, 1997.
- [6] P. Fernández-Cid and F.J. Casajús-Quirós. Enhanced quality and variety for chorus/flange units. In *Proc. of the first COST-G6 Workshop on Digital Audio Effects (DAFx)*, 1998.
- [7] Antti Huovilainen. Enhanced digital models for analog modulation effects. In *Proc. of the 8th Int. Conference on Digital Audio Effects (DAFx)*, 2005.
- [8] David T. Yeh, Jonathan S. Abel, and Julius O. Smith. Simplified, physically informed models of distortion and overdrive guitar pedals. In *Proc. of the 10th International Conference on Digital Audio Effects (DAFx)*, 2007.
- [9] David T. Yeh, Jonathan S. Abel, Andrei Vladimirescu, and Julius O. Smith. Numerical methods for simulation of guitar distortion circuits. *Computer Music Journal*, 32(2):23–42, 2008.
- [10] V. Verfaillie, U. Zölzer, and D. Arfib. Adaptive digital audio effects (a-dafx): A new class of sound transformations. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1817–1831, 2006.
- [11] Adam M. Stark, Matthew E.P. Davies, and Mark D. Plumbley. Rhythmic analysis for real-time audio effects. In *Proc. of the International Computer Music Conference (ICMC)*, 2008.
- [12] Alex Loscos and Thomas Aussenac. The wahwactor: a voice controlled wah-wah pedal. In *Proc. of the Conference on New Interfaces for Musical Expression (NIME)*, 2005.
- [13] Scott N. Levine. Effects processing on audio subband data. In *Proc. of the International Computer Music Conference (ICMC)*, 1996.
- [14] Florian Keiler, Daniel Arfib, and Udo Zölzer. Efficient linear prediction for digital audio effects. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFx)*, 2000.
- [15] Vincent Verfaillie and Philippe Depalle. Adaptive effects based on stft, using a source-filter model. In *Proc. of the 7th International Conference on Digital Audio Effects (DAFx)*, 2004.
- [16] Atsushi Marui and William L. Martens. Timbre of nonlinear distortion effects: Perceptual attributes beyond sharpness. In *Proc. of the Conference of Interdisciplinary Musicology (CIM)*, 2005.
- [17] William L. Martens and Atsushi Marui. Categories of perception for vibrato, flange, and stereo chorus: Mapping out the musically useful ranges of modulation rate and depth for delay-based effects. In *Proc. of the 9th International Conference on Digital Audio Effects (DAFx)*, 2006.
- [18] Vincent Verfaillie, Catherine Guastavino, and Caroline Traube. An interdisciplinary approach to audio effect classification. In *Proc. of the 9th International Conference on Digital Audio Effects (DAFx)*, 2006.
- [19] Perfecto Herrera-Boyer, Geoffroy Peeters, and Shlomo Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32:3–21, 2003.
- [20] Jakob Abeßer, Hanna Lukashevich, and Gerald Schuller. Feature-based extraction of plucking

- and expression styles of the electric bass guitar. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.
- [21] Adam Tindale, Ajay Kapur, George Tzanetakis, and Ichiro Fujinaga. Retrieval of percussion gestures using timbre classification techniques. In *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR)*, 2004.
- [22] Mauricio A. Loureiro, Hugo B. de Paula, and Hani C. Yehia. Timbre classification of a single musical instrument. In *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR)*, 2004.
- [23] Kerstin Dosenbach, Wolfgang Fohl, and Andreas Meisel. Identification of individual guitar sounds by support vector machines. In *Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx)*, 2008.
- [24] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report, IRCAM, Paris, 2004.
- [25] MarC Vinyes, Jordi Bonada, and Alex Loscos. Demixing commercial music productions via human-assisted time-frequency masking. In *Proc. of the AES 120th Convention*, 2006.
- [26] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005.
- [27] Simon Dixon. Onset detection revisited. In *Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx)*, 2006.
- [28] Wan-Chi Lee and C.-C. Jay Kuo. Musical onset detection based on adaptive linear prediction. In *Proc. of the International Conference on Multimedia & Expo (ICME)*, 2006.
- [29] Tae Hong Park. *Towards Automatic Musical Instrument Timbre Recognition*. PhD thesis, Princeton University, 2004.
- [30] Matti Karjalainen and Hanna Järveläinen. Is inharmonicity perceivable in the acoustic guitar? In *Proc. of Forum Acusticum 2005*, 2005.
- [31] Geoffroy Peeters and Xavier Rodet. Hierarchical gaussian tree with inertia ratio maximization for the classification of large musical instruments databases. In *Proc. of the 6th International Conference on Digital Audio Effects (DAFx)*, 2003.
- [32] Ethem Alpaydin. *Maschinelles Lernen*. Oldenbourg, München, 2008.
- [33] Hanna Lukashevich. Feature selection vs. feature space transformation in music genre classification framework. In *Proc. of the AES 126th Convention*, 2009.

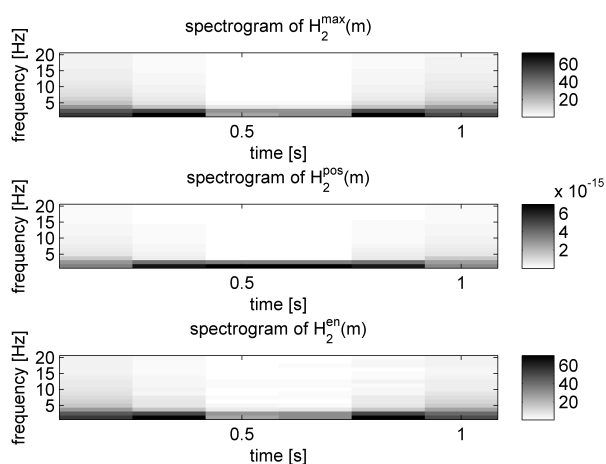


Fig. 3: Unprocessed Sample: In the spectrograms of the harmonic feature curves the energy is concentrated at the bottom, implying a relatively constant temporal course of the harmonics' frequency and magnitude.

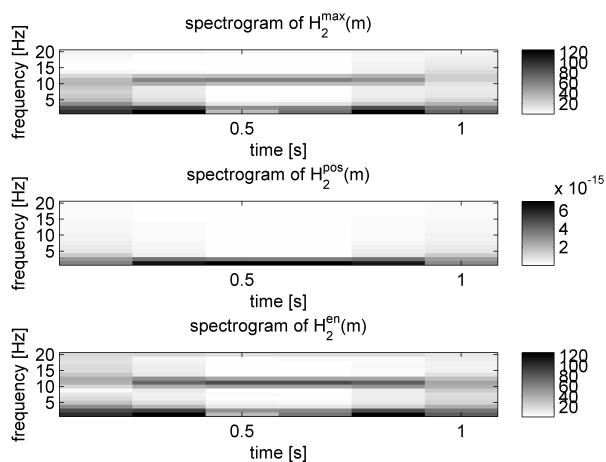


Fig. 4: Tremolo: The amplitude modulation caused by the effect can be seen clearly in the spectrograms of the harmonic feature curves $H_i^{max}(m)$ and $H_i^{en}(m)$. The peaks around 11 Hz indicate the modulation frequency.

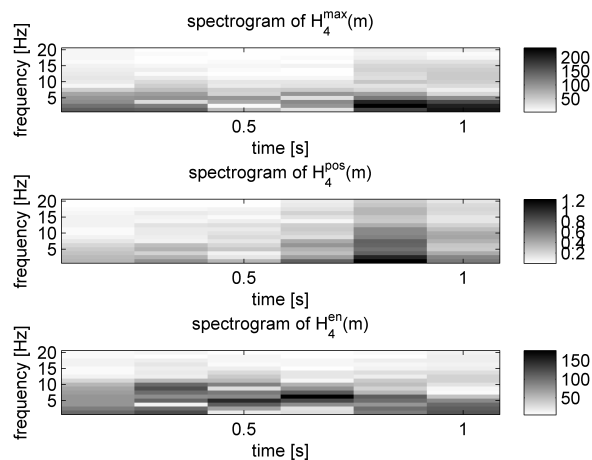


Fig. 5: Flanger: The distinct scatter of spectral energies is the result of the modulated notch frequencies of the comb filter. The continuous changes in frequency, magnitude and shape of the harmonics is reflected by the spectrograms of the harmonic feature curves.

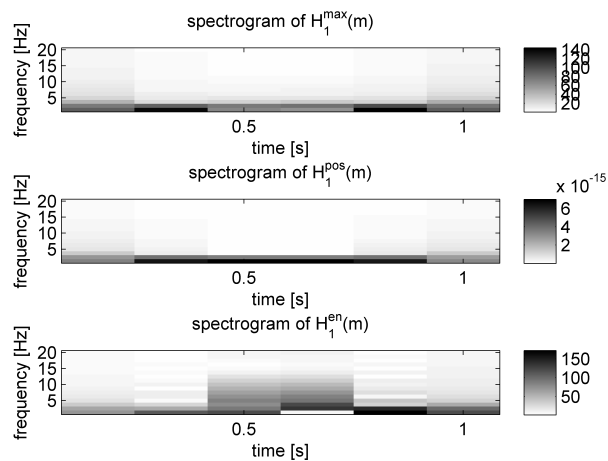


Fig. 6: Slapback Delay: The spontaneous rise of spectral width in the spectrogram of the harmonic feature curve related to the harmonic's energy $H_i^{en}(m)$ is the characteristic property of the slapback delay. The reason is the superposition of the delayed, energy-rich attack part of the sound with the sustain part which induces drastic changes to the shape of the harmonics.