AUTOMATIC DETECTION OF MULTIPLE, CASCADED AUDIO EFFECTS IN GUITAR RECORDINGS

Michael Stein,

Fraunhofer Institute of Digital Media Technology
Ilmenau, Germany
peter.michael.stein@hotmail.com

ABSTRACT

This paper presents a method to detect and distinguish single and multiple audio effects in monophonic electric guitar recordings. It is based on spectral analysis of audio segments located in the sustain part of 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. A novel database that consists of approx. 50000 guitar tones was assembled for the purpose of evaluation. Classification accuracy reached 99.2% for the detection and distinction of arbitrary combinations of six frequently used audio effects.

1. INTRODUCTION

Semantic music analysis is an active research topic, which aims to retrieve meaningful structural information about music data directly from the audio signal. This content-based meta data can be used for a variety of applications, including similarity-based music search and recommendation, interactive music games and enhanced music production tools. The utilized descriptors often focus on characterizing melodic, harmonic and rhythmic properties of musical pieces as well as their instrumentation. While this may be sufficient to characterize classical music, it misses one important facet of modern popular music: the usage of audio effects to alter the sound of single instruments and also effects processing of complete mixtures to put the finishing touch to them. Hence, effects processing can be considered an additional dimension of musical expression. It offers a broad range of possibilities, from subtle sound shaping to the creation of completely new sounds by feeding the instrument signal through a number of effects, each of them inducing a certain change to the sound. The electric guitar, which has been subject to effects processing for several decades now, has a special role in this context. In modern popular music it is often used as both lead and accompanying instrument and there are more than a few guitar players, whose signature sound is based on the usage of certain audio effects and their combination in particular.

Regarding semantic music analysis, we assume that existing techniques will benefit from the knowledge about the presence of audio effects applied to guitar sounds: 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. Using the a priori knowledge about the presence of these effects, one can try to remove them from the signal in a prior processing step or post process the error-prone output appropriately. In this paper, we show that automatic classification with Support Vector Machines based on extracted audio features provides a suitable approach for automatic detection of both single audio effects as well as multiple audio effects applied to guitar tones. Multiple effects in this context relates to cascaded single effects in the signal chain that alter the signal consecutively.

The remainder of this paper is organized as follows: After providing an overview of related work in Sec. 2, we present the individual processing stages of our approach in Sec. 3, introducing a novel audio feature extraction concept based on harmonic analysis of instrument sounds. In Sec. 4, we explain the performed experiments and discuss the obtained results. Finally, Sec. 5 concludes this work.

2. RELATED WORK

Audio effects are a multi-faceted research topic and various studies addressed topics such as the technical principles of audio effects and how these affect sound quality, emulation of analog circuitry behavior using digital signal processing or adaptive audio effects [1, 2, 3, 4]. However, in semantic music analysis they are scarcely addressed. Classification of musical instrument sounds is mainly performed on a level of instrument types or families, neglecting most of the timbral variations that can occur within the scope of a single instrument's sound, e.g. induced by playing styles or the usage of audio effects [5, 6]. One recent study investigated the detection of audio effects in recordings of electric guitar and bass but limited its scope to single effects [7]. We aim at extending this method to the detection of multiple audio effects applied to guitar sounds, which can be regarded a multi-label classification task, since a sound can be assigned a number of labels according to the number of effects applied to it [8]. The concept of multi-label classification has already been applied to a variety of tasks, including music genre classification and mood estimation [9, 10].

3. PROPOSED APPROACH

Within the great variety of different audio effects, a few of them can be considered as de facto standards for guitar sound shaping. In popular music, frequently used audio effects for sound processing comprise: *Nonlinear Processing* (NLP), *Modulation* (MOD) and *Ambience* (AMB) effects. Audio effects belonging to these three effect groups will be featured in almost every available multi effects device or in arrangements of single effects devices. For the experiments described in this paper, we chose two widely used audio effects from each group. These are: *Distortion* (DIS), *Overdrive* (OVD), *Chorus* (CHO), *Flanger* (FLA), *Feedback Delay* (FBD) and *Reverb* (REV). [1] provides detailed descriptions of

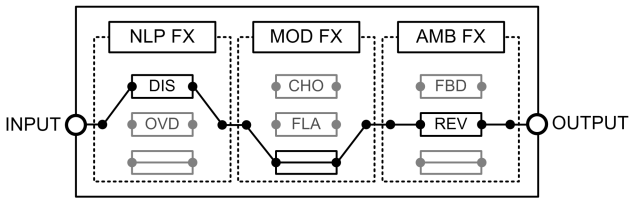

Figure 1: Schematic representation of a multi effects device containing the six audio effects investigated in this paper, grouped in three subsequent sections. In the depicted setup, two effects are activated (DIS and REV) while the modulation effects section is bypassed .

these and other audio effects. Although there exists no strict rule regarding the order of effects in the signal chain, most setups will feature an order comparable to that depicted in Figure 1. A potential user can choose to either activate a single effect or cascade effects of the different effect sections to achieve more complex sound alterations.

3.1. 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 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 effects applied to the signal. The proposed method consists of four stages: preprocessing, feature extraction, feature reduction and classification, which will be described in the following sections.

3.2. Preprocessing

Preprocessing aims to identify the starting time of the sustain part of a guitar tone. We compute the energy envelope of the tone and obtain an initial estimation of the end of the attack part with the fixed threshold method described in [11]. The actual transition between attack and sustain part of the tone 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 starting time behind the detected maximum.

3.3. Feature Extraction

An audio segment is first transformed into successive short time spectra using frames of 8192 samples, a hopsize of 512 samples and a Hann window at a sampling rate of 44.1 kHz. From these we extract a set of spectral, cepstral and harmonic features that has been recently employed for the detection of single audio effects [7]. Hence, only a brief overview of the feature extraction will be given in the following sections. The resulting feature vector has a total of 541 dimensions.

3.3.1. 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 [11, 12]. In addition, we calculate their first and second derivatives and apply highpass filtering to spectral centroid, roll-off and slope. To characterize the value range of the features, we use mean value and standard deviation. To capture their temporal progression, we calculate first four sample moments (namely mean value, variance, skewness and kurtosis) for each feature.

3.3.2. 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 contribute directly to the final feature vector along with their maximum value. In addition, we compute 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. We apply the same procedure to the standard deviations of the coefficients and repeat it for the first and second derivatives.

3.3.3. Harmonic Features

First, we estimate the fundamental frequency in every frame of the audio segment using autocorrelation and take the mode of all frames to obtain a more reliable estimation. Afterwards we determine the frequencies of individual harmonics by searching for the frequency bins with the highest magnitude in local ranges around integer multiples of the fundamental frequencies. This is necessary to account for inharmonicity, that is frequency deviations of the harmonics caused by the stiffness of the strings [13]. To reveal time-varying changes of the harmonics' frequencies, levels and shapes caused by the applied audio effect, they will not be represented by one single frequency bin but a narrow frequency band. Given the average frequency f_i of the i-th harmonic and a bandwidth Δk, we extract the following *Harmonic Feature Curves* H_i^* from the logarithmized, squared magnitude spectrogram X_{dB} of the audio segment:

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

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

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

$$\mathbf{k} = f_i - \Delta k, \ldots, f_i + \Delta k, \;\; \Delta k < f_0/2 \tag{4}$$

where m denotes the frame index, H_i^{max} the harmonic feature curve related to the maximum value, H_i^{pos} the harmonic feature curve related to the maximum position and H_i^{en} the harmonic feature curve related to the band energy of the i-th harmonic. $\mathbf{k}$ is a vector containing the frequency bin indices of the current analysis frequency band.

To derive figures that capture the sound alterations induced by the audio effects we perform short time spectral analysis on the harmonic feature curves of the first ten harmonics. From the resulting magnitude spectrograms we consider only the frequency range $0 \ldots 22$ Hz for analysis since this will contain the majority of low frequency modulations and variations. To characterize the spectral shape of the spectrograms, we extract the following

features frame-wise: mean value, variance and standard deviation, maximum value and position, spectral centroid, spread, skewness and kurtosis as well as the steady and the cumulated alternating component and their ratio. 72 figures per harmonic are derived by calculating mean value and standard deviation of the extracted features to characterize their value range.

To reduce the huge number of resulting figures we apply two grouping schemes. First, we average the figures over all harmonics. Secondly, we perform a tristimulus-like grouping driven by the concept, that the higher the order of the harmonics, they will be perceived rather as a group than as individual harmonics [12]. Figures related to the first harmonic remain unchanged, but we average the figures of the second to fourth as well as the figures of the fifth to tenth harmonic, thereby preserving valuable structural information.

Besides the spectral analysis of the harmonic feature curves, we also analyze the temporal progression of the harmonics' levels using the harmonic feature curves related to the band energy H_i^{en} (see Eq. 3). For unprocessed tones the levels will be decreasing slowly because the guitar string is performing a damped oscillation while the usage of audio effects can break this rule. We model an ideal progression of the harmonics' levels using linear regression of the band energy levels and compute the frame-wise differences to the real values 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 mean values and standard deviations of the differences and the absolute valued differences and apply the same grouping of figures as for the spectral analysis of harmonics. Furthermore, we calculate the ratio between the amounts of positive and negative valued differences and use statistical figures to evaluate the distribution of this ratio over the harmonics.

3.4. 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 a 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) [14] and Linear Discriminant Analysis (LDA) [15]. 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, smaller 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 (SVM) with a radial basis function kernel (RBF). More details on these methods can be found in [16].

There exist various strategies for classification of multi-label data. We utilize a method described in [8] that transforms the multi-label classification problem into one single-label classification problem by considering each different combination of single labels that exists in the data set as a single label. Table 1 illustrates this transformation for two effects. The main benefit of this approach is, that one only has to train one single-label classifier on the transformed data, whose predicted labels can be evaluated directly.

Item	DIS	CHO
1	x	
2		x
3	x	x

$\Longrightarrow$

It.	DIS	CHO	DIS & CHO
1	x		
2		x	
3			x

Table 1: Example of transforming multi-label data to single-label data using the example of two effects and their combination.

4. EXPERIMENTS AND RESULTS

4.1. Database

We used a database of recorded single guitar tones that were processed with different audio effects afterwards. Unprocessed tones as well as those processed with one single effect were taken from the *IDMT SMT Audio Effects* database, which is intended to be a public benchmark set for such tasks[1]. We extended this data set by processing the recorded tones with every of the 20 possible combination of two or three cascaded effects according to the setup depicted in Figure 1, which allows for 27 different effect combinations. Thereby, we used the same effect devices and settings as have been used for single effect processing to ensure that the distinction between multiple effects and the associated single effects will not be misleaded by differing effect parameter values. Furthermore, this resembles the performance of a real guitar player, who simply activates or deactives effects, most commonly with a foot control, without changing their settings, e.g. adding a bit of reverb to an already distorted guitar sound. We intend to extend the existing database with the newly created sounds[2]. In total, the database used for evaluation consists of 50544 tones, 1872 for each of the 27 possible effect combinations.

4.2. Experimental Setup

We evaluated the performance of the proposed approach with two experiments, which investigate how well single and multiple effects can be detected and distinguished from each other. In the first experiment we used the single effect labels, i.e. DIS, CHO, etc. as distinguishing criterion, whereas in the second one we used the effect group labels, i.e. NLP, MOD, etc. Accordingly, the two experiments were performed with 27 and 8 classes, respectively, always including the unprocessed samples as an additional class. For feature reduction we applied IRM with and without subsequent LDA as well as LDA solely. The number of features to be selected by the IRM algorithm was varied between 40 and 200 with a step size of 40 and the number of feature dimensions after LDA transform was set to 26 or 7. Regarding the SVM classifier, we varied the kernel parameter γ of the RBF kernel between 2^{-12} and 2^{4} and the cost parameter C between 2^{-4} and 2^{12} on a logarithmic scale. We randomly subdivided the database into train (75%) and test set (25%) and performed five-fold cross validation on the train set first to tune the parameters of feature reduction and classifier.

4.3. Results

Table 2 gives an overview of the results of the two experiments, depicting the best result obtained for each setup of the feature reduction stage. As shown there, the best mean classification accu-

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

[2] For further information please contact the author via email.

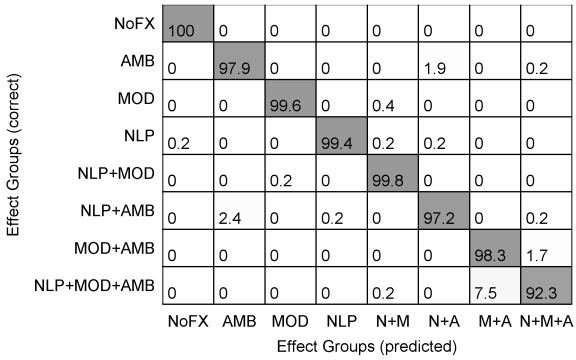

Figure 2: Confusion matrix [%] for the prediction of single and multiple effects on the effect group level. Mean classification accuracy is 98.1% and applied feature reduction is IRM with 160 features selected.

Effects		Effect Groups	
Feature Reduction	Result	Feature Reduction	Result
IRM 200	99.2	IRM 160	98.1
IRM 40 + LDA	97.4	IRM 200 + LDA	94.4
LDA	96.8	LDA	95.0

Table 2: Mean classification accuracies [%] of the two experiments for different setups of the feature reduction stage.

racy for the detection and distinction of single and multiple effects of 99.2% has been achieved by applying only IRM to the data. Applying LDA slightly decreases the results. The results for classification on the effect group level show a similar tendency. The best result of 98.1% was again achieved by applying only IRM to the data while applying LDA degraded the results once again. The complete confusion matrix for the best result of the second experiment is depicted in Figure 2. It shows that the majority of scatter results from incomplete detections of combinations of multiple effect , most notable for the combination of all three effect groups.

5. CONCLUSIONS

In this paper we presented a machine learning approach for simultaneous detection and distinction of single and multiple audio effects in monophonic electric guitar recordings. We showed that a method, designed for the detection of single audio effects, could successfully be adapted to detect and distinguish arbitrary combinations of up to three audio effects. A novel database of isolated guitar tones was assembled for evaluation purpose. It is intended as an open benchmark for the given and related tasks. The obtained results indicate that the proposed approach might be a valuable enhancement for existing music analysis systems. Future steps will focus on modifying the proposed method with regard to detection of single and multiple effects with a priori knowledge of single effects only, to improve its scalability.

6. REFERENCES

[1] Udo Zölzer, Ed., *DAFX - Digital Audio Effects*, John Wiley & Sons, Chichester, 2002.

[2] Antti Huovilainen, "Enhanced digital models for analog modulation effects," in *Proc. of the 8th Int. Conference on Digital Audio Effects (DAFx)*, 2005.

[3] David T. Yeh, Jonathan S. Abel, Andrei Vladimirescu, and Julius O. Smith, "Numerical methods for simulation of guitar distortion circuits," *Computer Music Journal*, vol. 32, no. 2, pp. 23–42, 2008.

[4] V. Verfaille, 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*, vol. 14, no. 5, pp. 1817–1831, 2006.

[5] Perfecto Herrera-Boyer, Geoffroy Peeters, and Shlomo Dubnov, "Automatic classification of musical instrument sounds," *Journal of New Music Research*, vol. 32, pp. 3–21, 2003.

[6] 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.

[7] Michael Stein, Jakob Abeßer, Christian Dittmar, and Gerald Schuller, "Automatic detection of audio effects in guitar and bass recordings," in *Proc. of the AES 128th Convention*, 2010.

[8] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, pp. 1–13, 2007.

[9] Hanna Lukashevich, Jakob Abeßer, Christian Dittmar, and Holger Grossmann, "From multi-labeling to multi-domain-labeling: A novel two-dimensional approach to music genre classification," in *Proc. of the 10th International Conference on Music Information Retrieval (ISMIR)*, 2009.

[10] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multilabel classification of music into emotions," in *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR)*, 2008.

[11] Geoffroy Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," Tech. Rep., IRCAM, Paris, 2004.

[12] Tae Hong Park, *Towards Automatic Musical Instrument Timbre Recognition*, Ph.D. thesis, Princeton University, 2004.

[13] Matti Karjalainen and Hanna Järveläinen, "Is inharmonicity perceivable in the acoustic guitar?," in *Proc. of Forum Acusticum 2005*, 2005.

[14] 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.

[15] Ethem Alpaydin, *Maschinelles Lernen*, Oldenbourg, München, 2008.

[16] Hanna Lukashevich, "Feature selection vs. feature space transformation in music genre classification framework," in *Proc. of the AES 126th Convention*, 2009.