PARAMETER EXTRACTION FOR BASS GUITAR SOUND MODELS INCLUDING PLAYING STYLES

Gerald Schuller

Ilmenau University of Technology
Department for Media Technology
98693 Ilmenau, Germany

Jakob Abeßer, Christian Kehling

Fraunhofer IDMT
Semantic Music Technologies Group
98693 Ilmenau, Germany

ABSTRACT

We present a system to realistically model the sound of bass guitars, and how to estimate the corresponding parameters from the sound of a bass guitar alone, without other physical measurements. Our model includes plucking and expression styles of the musician, like vibrato or bending, and the string number for a realistic modeling and reproduction of the sound. We show that we can estimate the playing techniques and the string number with relatively high accuracy.

Index Terms— instrument coder, sound model, bass guitar, playing styles

1. INTRODUCTION

Application scenario: We have a musical piece with a mix of instruments, and would like to analyze, transmit, or modify it. In the first processing step we have a hypothetical system for separating out each instrument from the mix. In the next step, we analyze the sound of each instrument, estimating parameters for a model of the instrument sound. This can be seen as an encoding system. In the decoding step, we take these parameters and feed them into a sound synthesis algorithm to regenerate a sound as close as possible to the original sound or at least natural and in a way that captures the most important perceptual characteristics of the instrument. On this decoding side, all the sounds from each instrument model can be put back together in a mix. But now we have the additional advantage that we can attenuate or emphasize each instrument individually if we wish, or we can generate a surround mix where we can modify the position of each instrument sound in space. A specific advantage of the model generated sound is that we can also modify the characteristics of each instrument sound, for instance modifying the instrument or changing playing styles of the instrument. These might be important possibilities for future multimedia systems. A similar interactive scenario is described in [1].

The goal of this paper is to develop an algorithm to transcribe electric bass guitar recordings and estimate the most important perceptual parameters for each played note. These parameters can be fed to a sound synthesis model in the decoding step in order to recreate the original instrument sound [2]. All required parameters are estimated from isolated bass guitar recordings without any additional physical measurements of the instrument itself.

2. PROBLEM TO SOLVE

We need to model the following effects for a natural and similar reproduced sound. The most influential parts of an electric guitar are the strings, the magnetic pick-up, and the passive electrical tone control. Body resonances only have a minor influence on the resulting tone and will not be taken into account here. The guitar strings determine the basic sound since when vibrating, they are the primary sound source.

The sound is mainly affected by the string material, tension, and stiffness. These features manifest primarily in frequency shifts of partial vibrations also known as the effect of inharmonicity [3]. Electromagnetic pick-ups capture the string vibration depending on their position on the instrument neck and the corresponding possible displacement of partials.

Another important means for the musician to manipulate the tone are the plucking and expression styles, which are used to play single notes on the instrument. In this work we distinguish 5 different plucking styles—finger style, picked, muted, slap pluck, and slap thumb—as well as 6 different expression styles—bending, slide, vibrato, harmonics, and dead notes—executed with the fingering hand in addition to non-decorated, normal expression style. See [4] for a detailed description of the playing techniques. In this publication, we focus on a precise description of electric bass guitar tracks. An extension of the proposed methods towards the analysis of the electric guitar was presented in [5].

3. RELATED WORK

The *bass line* is considered to be the dominant melodic voice in the lower pitch register with fundamental frequencies between around 40 Hz and 400 Hz. Due to the low fundamental frequency range of bass notes, *downsampling* is commonly

applied to the analyzed audio signal to accelerate the transcription process [6, 7, 8]. At the same time, harmonic components from other instruments in higher frequency ranges are filtered out. Ryynänen and Klapuri estimate a variable, context-dependent upper f_0-limit for the bass line [9]. In other publications, signal components or the percussion instruments [10] are removed in the spectrum before the bass line is transcribed by applying different source separation techniques such as the harmonic/percussion sound separation (HPSS) algorithm. Spectral whitening can be applied to make the transcription algorithm more robust to different timbres of the bass instrument [9]. *Note detection* is performed either in the time domain [6] by envelope extraction methods or in the frequency domain—usually after several frame-wise f_0 estimates are grouped to note events [11, 9, 7].

Due to its computational efficiency, the Short-time Fourier Transformation (STFT) is the most often used *spectral estimation* method [6, 11, 9]. Other spectral representations such as the instantaneous frequency (IF) spectrogram [6, 7] or the constant-Q spectrogram [12] are computed to improve the achievable frequency resolution in the lower frequency bands. Ryynänen and Klapuri present a hybrid transcription framework for bass and melody transcription in polyphonic music [9] by combining a Hidden Markov Model (HMM) with two modeling strategies—acoustic note modeling and a musicological model of the most probable note transitions.

4. NEW APPROACH

4.1. Development data sets

Two development sets *DS-1* and *DS-2* were taken from the *IDMT-SMT-Bass* dataset (previously published in [4]) and used for parameter optimization. The development set *DS-1* comprises 550 randomly selected isolated bass guitar notes (50 notes for each plucking and expression technique) and the development *DS-2* comprises 1711 notes, which were recorded with the same instrument (Fame Baphomet 4 NTB bass guitar) that was also used to record the bass lines in the evaluation dataset introduced in Section 5.1. In the following sections, the individual processing steps of the proposed transcription algorithm are presented in detail.

4.2. Pre-processing & Spectral Estimation

First, we convert the audio signal to a monaural signal if necessary and down-sample to a sampling frequency of $f_s \approx 5.51$ kHz. Then, two different spectral representations are extracted. First, a *Short-time Fourier Transform (STFT) spectrogram* X is computed using a blocksize of 512 samples and hopsize of 32 samples. The STFT spectrogram is used for the envelope modeling as will be explained in Section 4.5. Second, a *reassigned spectrogram* X_{IF} based on the instantaneous frequency (IF) is computed with the same blocksize and hopsize values. We estimate the instantaneous frequency based on the method proposed by Abe et al. in [13]. We use a logarithmic frequency axis with a fine resolution of 120 bins per octave in the range between 29.1Hz and $f_s/2$. In each time frame t, the magnitude values of the STFT spectrogram are reassigned and accumulated towards the logarithmic frequency bins that correspond to the IF values at the original frequency positions. Since sinusoidal peaks tend to produce stable IF values in the surrounding frequency bins, sharper peaks can be seen at frequency positions of the sinusoidal signal components in the IF spectrogram. The mapping from the continuous IF to the discrete logarithmic frequency scale is performed in order to perform the cross-correlation with a harmonic comb filter as will be explained in Section 4.4.

4.3. Onset Detection

To detect the note onset time, we propose a novel onset detection function that measures the *harmonic novelty*. The basic idea is to detect signal parts, where harmonic components (fundamental frequency and overtones) begin. These components show a sparse energy distribution over the frequency with magnitude peaks approximately at multiples of the fundamental frequency of each note. Hence, we design a matched filter in the time-frequency domain in such way, that it resembles a rise in magnitude across time and a sparse peak across frequency. We compute the two-dimensional cross correlation between the IF spectrogram and the matched filter. For details, please refer to [14]. By summing across frequency, we obtain an onset detection function $o(t)$. Onsets t_{on} are then detected at all local maxima of $o(t)$ greater than $o_{\text{min}} = 0.2 \max_t o(t)$.[1] This empirical threshold was found based on the development set *DS-1* with manually annotated onset positions by maximizing the F-measure ($FM = 0.95$).

4.4. f_0-tracking & Offset Detection

Two processing steps are performed to track the fundamental frequency of each note over time: pre-estimation of the note's f_0 and f_0 tracking. First, for the n-th note, the spectral frames in X_{IF} are averaged over the first 20 % of the time frames between the onset positions $t_{\text{on}}(n)$ and $t_{\text{on}}(n+1)$ to obtain an *accumulated spectrum* $X_{\text{IF,acc},n}(f)$. We focus on the beginning period to prevent a smearing of harmonic peaks for notes played with modulation techniques such as bending, vibrato, and slides, which have a time-varying fundamental frequency. The pre-estimate $\hat{f}_{0,n}$ is detected at the frequency bin with the highest cross-correlation between $X_{\text{IF,acc},n}(f)$ and a *harmonic comb filter* $c(f)$, which has combs at the harmonic frequency positions

$$f_k \approx f_0 \cdot (k+1)\sqrt{1+\beta\cdot(k+1)^2} \qquad (1)$$

[1] However, this approach of using a fixed threshold could lead to missed note events for recordings with a large dynamic range.

[15] on a logarithmic frequency axis (as used before for $X_{\text{IF},n}$). The inharmonicity coefficient was set to $\beta = 3E-4$, which is an average over multiple notes from the set *DS-1*. Using 500 notes from the development set *DS-1* [2], we compared comb filters with a varying number of harmonic peaks for the task of pitch detection. Furthermore, we compared comb filters with peaks having unit magnitudes and comb filters with doubled magnitude on the first two peaks. We achieved the highest detection accuracy of 0.98 (percentage of correctly identified note pitches) for a comb filter with 10 combs and an emphasis of the magnitudes of the first two combs as shown in Figure 1. We did not observe an improvement in pitch detection accuracy by using linearly decaying values for the filter peaks.

The frame-wise f_0-tracking is initialized at the frame t_{Start}, which was chosen to be in the middle of the period used for averaging the spectrogram (as explained above). The tracking is performed over adjacent frames in two directions—backwards until reaching the note onset and forwards until reaching the following onset or the last frame. We use a *continuity-constraint*, i.e., in each frame, we only consider the frequency bins around the f_0 bin from the preceding frame as potential f_0 candidates. Again, the highest cross-correlation between the spectral frame and the comb filter is retrieved.

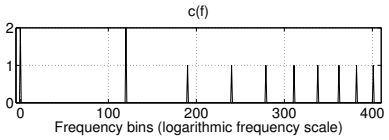

Fig. 1. Optimal harmonic comb filter $c(f)$ used for f_0-tracking based on a logarithmic frequency axis.

The maximum cross-correlation value is stored for each frame—high values indicate a harmonic magnitude characteristic of the spectrum, low values indicate a percussive, wide-band characteristic. We determine the *offset position* $t_{\text{off}}(n)$ of each note, where the maximum cross-correlation value remains below a threshold of 0.05 for at least 4 frames or a new note begins.

4.5. Spectral Envelope Modeling

Here we model the spectral magnitude envelope of a given note using a simple parametric model. We focus on the fundamental frequency and the overtones and neglect wide-band noise-like signal components such as the attack transients. The main motivation is to parametrize all possible spectral envelopes of bass guitar notes using a simple model. Each time frame in the STFT magnitude spectrum $X(f,t)$ is modeled as a sum of magnitude-scaled atom functions $h_X(f)$ shifted in frequency, which represent the harmonic components:

$$X(f,t) \approx \sum_{k=0}^{N_{\text{Harm}}} a_k(t)\; h_X\left(f - f_k(t)\right) \qquad (2)$$

The time-varying harmonic magnitudes are denoted as $a_k(t)$. The atom function $h_X(f)$ is the Fourier transform of the Hanning window $h(t)$, which is applied in the time domain to compute the STFT spectrogram X.

We initially estimate the inharmonicity coefficient β in (1) at t_{Start} in the beginning of the note decay part (compare Section 4.4) and assume it to be constant over the duration of the note. Therefore, we perform a grid search within $\hat{\beta} \in [0, 0.001]$. For each estimate of $\hat{\beta}$, we compute the hypothetical spectrum using (1). We use an optimization algorithm to find the set of coefficients a_k and β which best describe the observed spectrum (see [14] for more details). After the envelope modeling, each note is described by a set of envelope parameters $[a_k(t), \beta, f_0(t)]$, which are then used for feature extraction as will be described in the following sections. In Figure 2, an example of the note modeling is shown for a bass guitar note with vibrato. The magnitude envelopes of the overtones were well-captured (including the typical phenomena of *string beating* [15]), attack transients were neglected due to the discussed modeling approach.

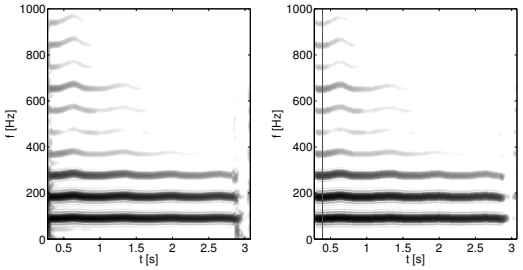

Fig. 2. STFT spectrogram $X(f,t)$ of a vibrato note: original (left) and modeled (right). The start frame for the optimization is shown as vertical line in the right figure.

4.6. Estimation of Plucking Style, Expression Style, and String Number

In order to estimate the *instrument-level parameters* plucking style, expression style, and string number, we extract various audio features and train a statistical classification model for each parameter based on given training examples from the development set *DS-2*. For instance, we compute the spectral centroid, harmonic magnitude slope, or inharmonicity factor as features. The full set of feature is detailed in [16, 4, 5].

For each classifier, we first normalize the feature values to zero mean and unit variance. Second, the supervised feature selection method Inertia Ratio Maximization using Feature Space Projection (IRMFSP) [17], which takes the class

[2] The 50 notes played with the *dead-note* expression style were excluded since they are percussive without a perceivable stable pitch

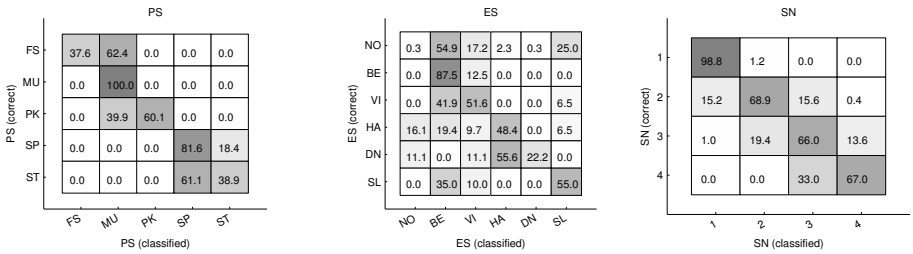

Fig. 3. Confusion matrices for the estimation of instrument-level parameters. All values given in percent.

labels into account, is applied to reduce the dimensionality of the feature space to $D = 60$. We used features as described and listed in [5]. Third, the feature space transformation method Linear Discriminant Analysis (LDA) is applied to further reduce the dimensionality of the feature space to $D = N_{\text{classes}} - 1$. Finally, a Support Vector Machine (SVM) classifier with a radial basis function (RBF) kernel is trained for each of the three classification tasks. The fret number is derived depending on the expression style, three cases are differentiated. For *dead-notes*, the fret number is not considered to be relevant and the string number is set to the string number of the closest note, which was played in one of the expression styles *normal*, *vibrato*, *bending*, or *slide* (this makes the bass line easier to play). Since *harmonics* with a given mode can be played on multiple fret positions, we set the fret number to be preferably close to the fret numbers of previous notes based on the estimated mode $\hat{m}$. The fret number is derived from the pitch and the string number as explained in [16].

5. EVALUATION

5.1. Dataset

The evaluation of the proposed methods is performed under idealized conditions. The previously published *IDMT-SMT-BASS-SINGLE-TRACKS*[3] dataset is used for evaluation. It consists of 17 bass lines that cover different music styles (blues, rock, funk, bossa nova, and hip hop). The bass lines consist of around 1000 notes and cover all discussed plucking and expression styles as well as all 4 strings of the bass guitar.

For brevity we only present the results for the plucking and expression styles and string number. More detail can be found for instance in [18].

5.2. Instrument-level Evaluation

In order to evaluate the estimation of the instrument-related parameters plucking style (PS), expression style (ES), and string number (SN), three classification experiments were performed as follows. In order to eliminate the onset and pitch estimation as potential error sources, we use ground truth annotations for the note pitch, onset, and offset instead. The three classifiers are trained with notes from the development set *DS-2*.

The confusion matrices for the estimation of the instrument-related parameters PS, ES, and SN are shown in Figure 3. For PS and SN classification, a main diagonal is clearly visible—mean classification accuracy values of 0.64 and 0.75 were achieved. For the ES classification, only the BE class shows satisfying results. The other classes—especially NO and DN show strong confusions towards other classes. The mean accuracy for ES classification is 0.44. For the plucking style classification, we observe confusions between the fingerstyle (FS) and muted (MU) class, which only differ in the amount of damping while resemble each other with respect to the spectral envelope. Also the confusion between the two slap classes slap-pluck (SP) and slap-thumb (ST) seems reasonable due to their similar sound production on the instrument. The observed confusions between NO and BE might result from small f_0 fluctuation for regular notes without modulation (NO), which can be confused as bending notes (BE). Also, percussive dead-notes (DN) sometimes show a very short presence of a fundamental frequency and overtones, this might explain the confusions towards the harmonics (HA) class.

To the best knowledge of the authors, only in [5], a system for estimating the same parameters from electric guitar recordings was presented. The authors achieved slightly better accuracy values for string number (0.82 for 6 classes), plucking style (0.93 for 3 classes), and expression style (0.83 for 6 classes), but based on an additional filtering of non-plausible classification results.

6. CONCLUSIONS

We presented approaches to automatically transcribe and estimate parameters from isolated bass guitar recordings for a realistic model of bass guitar sounds. We found that the plucking and expression style of the musician playing the instrument are an important part of the character of the sound. Our experiments showed that these parameters can be estimated from isolated instrument recordings with relatively high accuracy.

[3] see `http://www.idmt.fraunhofer.de/en/Departments_and_Groups/smt/bass_lines.html`

7. REFERENCES

[1] J. Inseon, P. Kudumakis, M. Sandler, and K. Kyeongok, "The MPEG Interactive Music Application Format Standard," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 150–154, 2011.

[2] Jakob Abeßer, Patrick Kramer, Christian Dittmar, and Gerald Schuller, "Parametric Audio Coding of Bass Guitar Recordings using a Tuned Physical Modeling Algorithm," in *Proc. of the 16th International Conference on Digital Audio Effects (DAFx-13)*, Maynooth, Ireland, 2013.

[3] Isabel Barbancho, Senior Member, Lorenzo J. Tardón, Simone Sammartino, and Ana M. Barbancho, "Inharmonicity-Based Method for the Automatic Generation of Guitar Tablature," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1857–1868, 2012.

[4] 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)*, Dallas, USA, 2010, pp. 2290–2293.

[5] Christian Kehling, Jakob Abeßer, Christian Dittmar, and Gerald Schuller, "Automatic Tablature Transcription of Electric Guitar Recordings by Estimation of Score- and Instrument-Related Parameters," in *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14), Erlangen, Germany, September 1-5*, 2014.

[6] Christian Dittmar, Karin Dressler, and Katja Rosenbauer, "A Toolbox for Automatic Transcription of Polyphonic Music," *Proceeding of the Audio Mostly Conference*, pp. 58–65, 2007.

[7] Masataka Goto, "A Real-Time Music-Scene-Description System - Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-World Audio Signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, Sept. 2004.

[8] Justin Salamon and Emilia Gómez, "A Chroma-Based Salience Function for Melody and Bass Line Estimation From Music Audio Signals," in *Proc. of the 6th Sound and Music Computing Conference (SMC)*, Porto, Portugal, 2009, pp. 23–25.

[9] Matti P. Ryynänen and Anssi Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, pp. 72–86, Apr. 2008.

[10] Emiru Tsunoo, Nobutaka Ono, and Shigeki Sagayama, "Musical Bass-Line Pattern Clustering and its Application to Audio Genre Classification," in *Proc. of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, Kobe, Japan, 2009, pp. 219–224.

[11] Matti Ryynänen and Anssi Klapuri, "Automatic Bass Line Transcription from Streaming Polyphonic Audio," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP')*, Honolulu, Hawaii, USA, 2007, pp. 1437–1440.

[12] Emiru Tsunoo, George Tzanetakis, Nobutaka Ono, and Shigeki Sagayama, "Beyond Timbral Statistics: Improving Music Classification Using Percussive Patterns and Bass Lines," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 1003–1014, 2011.

[13] Toshihiko Abe, Takao Kobayashi, and Satoshi Imai, "Robust Pitch Estimation with Harmonics Enhancement in Noisy Environments based on Instantaneous Frequency," in *Proc. of the 4th International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, PA, USA, 1996.

[14] Jakob Abeßer, *Automatic Transcription of Bass Guitar Tracks applied for Music Genre Classification and Sound Synthesis*, Ph.D. thesis, Ilmenau University of Technology, Ilmenau, Germany, 2014.

[15] Neville H. Fletcher and Thomas D. Rossing, *The Physics Of Musical Instruments*, Springer, New York, London, 2nd edition, 1998.

[16] Jakob Abeßer, "Automatic String Detection for Bass Guitar and Electric Guitar," in *From Sounds to Music and Emotions - 9th International Symposium, CMMR 2012, London, UK, June 19-22, 2012, Revised Selected Papers*, Mitsuko Aramaki, Mathieu Barthet, Richard Kronland-Martinet, and Sølvi Ystad, Eds., London, UK, 2013, vol. 7900, pp. 333–352, Springer.

[17] 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)*, London, UK, 2003, pp. 1–6.

[18] Jakob Abeßer and Gerald Schuller, "Instrument-centered Music Transcription of Bass Guitar Tracks," in *Proc. of the AES 53rd Conference on Semantic Audio*, London, UK, 2014.