MODELING MUSICAL ATTRIBUTES TO CHARACTERIZE ENSEMBLE RECORDINGS USING RHYTHMIC AUDIO FEATURES

Jakob Abeßer[*] [○] *Olivier Lartillot*[⋆] *Christian Dittmar*[○] *Tuomas Eerola*[⋆] *Gerald Schuller*[○]

[○] Fraunhofer IDMT, Ilmenau, Germany
[⋆] Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland

ABSTRACT

In this paper, we present the results of a pre-study on music performance analysis of ensemble music. Our aim is to implement a music classification system for the description of live recordings, for instance to help musicologist and musicians to analyze improvised ensemble performances. The main problem we deal with is the extraction of a suitable set of audio features from the recorded instrument tracks. Our approach is to extract rhythm-related audio features and to apply them for regression-based modeling of eight more general musical attributes. The model based on Partial Least-Squares Regression without preceding Principal Component Analysis performed best for all of the eight attributes.

Index Terms— Music performance analysis, improvisation, micro-timing, regression analysis, onset detection

1. INTRODUCTION

Improvisation and ensemble play are common techniques in modern music genres like swing, blues, or funk, but have rarely been studied in the Music Information Retrieval literature. In this paper, we investigate trio ensembles including bass guitar, electric guitar, and drums. Professional multi-track recordings allow the separate analysis of all instrument tracks in order to omit the error-prone step of source separation. This paper is organized as follows. We outline the goals of this publication in Sect. 2 and give a brief overview over related work in Sect. 3. In Sect. 4, we explain the music recording and data annotation process. Furthermore, we illustrate the extracted audio features and the regression analysis configurations that we evaluated. Finally, we discuss the results in Sect. 5 and give a conclusion in Sect. 6.

2. GOALS

In this work, we aim to develop a estimation model that allows to automatically characterize music ensemble performances in terms of different attributes. In order to facilitate interpretations, we want to model eight rhythmic attributes that correspond to commonly used musical expressions. Based on rhythm-related audio features, we aim to evaluate three different configurations of regression analysis models and preceding feature space reduction in order to estimate values of these attributes for unknown music excerpts. This automatic approach allows to analyze large data sets of recorded live performances and could stimulate further research on the performance characteristics of different musicians.

3. PREVIOUS APPROACHES

An overview over different approaches related to music performance analysis is given by Widmer and Goebl in [1]. Most authors focus on characterizing the performance of one or multiple musicians based on different rhythmic properties. The impact of timing in drummers' performance was covered among others in [2]. A study concerning the perception of groove has been previously presented in [3], the role of groove in jazz trios was investigated in [4]. In order to describe musical properties automatically, the extraction of audio features is an indispensable step. *Pulse clarity* was modeled before in [6] by means of different rhythmic audio features. Eerola et. al. used a similar approach based on regression analysis in order to use audio features to model different emotions in music [7].

4. NEW APPROACH

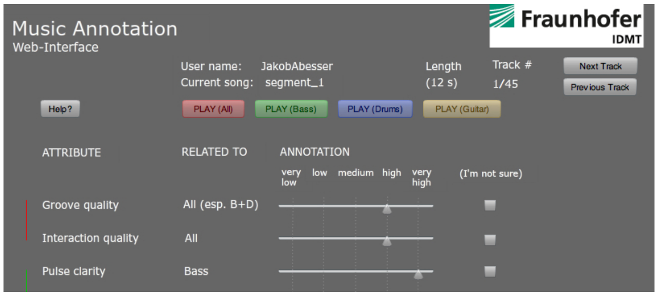
4.1. Recording

The music recordings were performed under supervision of an audio engineer in the music studio of the University of Jyväskylä in the beginning of June 2010. Three groups, each with three proficient musicians (electric guitar, bass guitar, drums) were recruited from the local music conservatory. Each group was asked to record one musical performance within each of the genres blues, funk, and swing each with at least one improvised solo part from each of the three musicians. Overall, 72 minutes of multi-track audio consisting of

[*] The first author performed parts of the work between May and August 2010 during a research stay at the Department of Music, University of Jyväskylä. All correspondence shall be addressed to abr@idmt.fraunhofer.de

isolated audio tracks for bass guitar, guitar, bass drum (drum-set), snare-drum (drum-set) and hi-hat / cymbals (drum-set) were recorded.

4.2. Annotation



User name: JakobAbesser
Current song: segment_1
Length (12 s)
Track # 1/45

ATTRIBUTE	RELATED TO	ANNOTATION
Groove quality	All (esp. B+D)	high
Interaction quality	All	high
Pulse clarity	Bass	very high

Fig. 1. Section of the annotation interface (screenshot)

In order to characterize improvised ensemble recordings, we performed preliminary interviews with some of the participating musicians. Here, we could identify 8 musical attributes to focus on in our experiments. *Groove quality* is mainly related to the perceived performance quality of the rhythm section (bass and drums). *Interplay quality* refers to the performance quality of the complete trio. In addition to these general attributes, we use *pulse clarity* and *predictability of rhythm* for the bass track, the drum track, and the guitar track individually. *Pulse clarity* describes how well general rhythmic properties such as tempo, time signature, and beat positions can be perceived by just listening to the bass track, the guitar track, or the drum track in isolation. We expect the attribute *predictability of rhythm* to take higher values in segments, where a particular instrument plays a constant repeating rhythm. 8 attributes are used in total.

We selected 45 excerpts from the recordings of 5 to 12 seconds duration. These excerpts were taken from different segment types as for instance chorus, bass solo, drum solo, and guitar solo. For the annotation process, a web-interface based on Flash was developed. In Fig. 1, a section of the interface is depicted. Both the annotation and user management data were stored in an SQL database. 17 participants provided a full set of 45×8 attribute annotations. We computed the Cronbach alpha, which measures the internal consistency and hence reliability of the annotation data. The Cronbach alpha was > 0.76 for all attributes after the annotation and > 0.84 after an outlier detection based on the inter-subject cross correlation (ISC). One annotation was removed per attribute.

4.3. Onset Detection

The detection of temporal note events is a crucial step for a preceding extraction of rhythmic audio features. Therefore, we utilized the 5 different *onset-classes snare drum & rimshot* (**SR**), *bass drum* (**BD**), and *cymbals & hi-hat* (**CH**) (related to the drum track) as well as *bass* (**BS**) and *guitar* (**GT**) related to the bass guitar and the guitar track. We used the drum transcription algorithm previously presented in [8] to detect all events related to the onset classes **SR**, **BD**, and **CH**.

For the onset detection in the bass track (**BS**) and the guitar track (**GT**), we used the *mironsets* function with 'Log' option in MIRtoolbox 1.3.2 [1]. Here, an amplitude envelope is extracted through the computation of a power spectrogram (frame size 100 ms, hop factor 10%, Hanning window), which is subsequently accumulated along frequencies. Prominent peaks are extracted from the common logarithm (base 10) of the resulting envelope curve using an adaptive peak picking algorithm that selects local maxima whose distances with their previous and successive local minima are both higher than 1% of the total amplitude range of the onset curve.

4.4. Feature Extraction

For each segment, the time position and total number of beats have been manually annotated since different automatic beat-tracking systems such as BeatRoot [2] did not lead to satisfying result for the analyzed data. After the automatic onset detection, explained in the previous section, all detected note onset values are given in seconds. We use the beat positions to map the onset times to multiples of measure lengths. This musical time representation is tempo-independent, thus it is better suited for the extraction of rhythmic features. The tempo of musical performances rarely remains constant [2]. We compute local tempo values by analyzing the time difference between adjacent beats positions. We take both the mean and standard deviation over all measures as features to characterize the tempo variations during the performance.

To measure the note density of each instrument track over time, we follow two approaches as previously presented in [9]. All onset-related features explained hereafter are computed for each onset class separately. If not stated otherwise, we derive features by computing the mean and standard deviation over all feature values and over the whole segments. First, we compute the number of notes per measure for each onset class. Second, we derive the inter-onset interval (IOI) expressed as fractions of the measure length. The IOI measures the onset distance between adjacent notes, a small average IOI value indicates a high note density.

Improvised music is often characterized by small onset deviations between notes played by an instrument and the actual beat positions. These are referred to as *micro-timing*. Our approach is to investigate four different rhythmic subdivisions $q \in \{4, 8, 16, 32\}$. For each subdivision, all measures are divided into q equidistant beats each. Based on that, we perform a *quantization* of all detected note onsets. This means that each note is mapped to its closest adjacent beat by shifting its onset value. We store this shift for each note and divide

[1] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

[2] http://www.eecs.qmul.ac.uk/$\sim$simond/beatroot/

it by the beat distance to obtain a relative distance measure in percent. In order to measure the micro-timing within each instrument track, we average the onset shift values over all notes for a given subdivision q and compute the mean and standard deviation. This is done separately for each onset class. These features allow to measure whether the notes of an instrument tend to be played shortly after or ahead of the beat.

In order to investigate whether the rhythmic structure of an instrument track is often varied or whether it is constantly repeated, we proceed as follows. After mapping all note onset values to an eighth-note subdivision ($q = 8$), we segment each instrument track into note sequences of one measure length. Then, we use the *Levensthein distance* d_L to compute the rhythmic similarity between the note sequences in two arbitrary measures (see [9] for more details).

We compute the matrix $\boldsymbol{S} \in \mathbb{R}^{N \times N}$ ($s_{i,j} \in [0, 1]$) with N denoting the number of measures and $s_{i,j}$ denoting the similarity between measures i and j. To characterize the average rhythmic similarity and the degree of variation over all measures, we compute the mean and standard deviation over all non-diagonal elements of $\boldsymbol{S}$ as features. To obtain a local similarity measure, we investigate the similarity between consecutive measures in an instrument track. Therefore, we concatenate all elements $s_{i,j}$ with $|i - j| = 1$ below the main diagonal of $\boldsymbol{S}$ in a vector $\boldsymbol{s_c} \in \mathbb{R}^{N-1}$. To measure the rhythmic deviation between consecutive measures, we compute the difference vector $\boldsymbol{\Delta s_c}$ between adjacent elements of $\boldsymbol{s_c}$ and compute the mean and standard deviation over all elements of $\boldsymbol{\Delta s_c}$ as features.

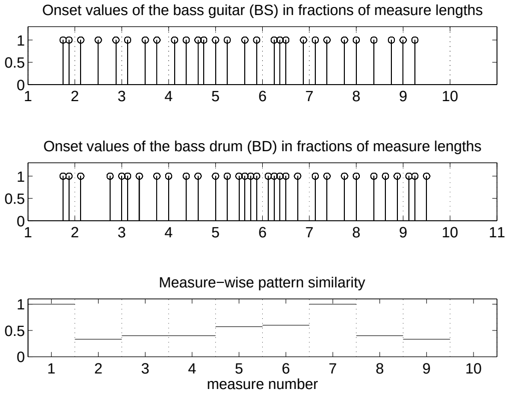


Fig. 2. Rhythmic similarity between different instrument tracks.

In groove-based ensemble music, different instrument often play in the same rhythm. To detect sequences with equal onset sequences between instrument tracks, we use the same distance measure as described above to measure the similarity between two simultaneously performed instrument tracks. In Fig. 2, 10 consecutive measures of the bass guitar track and the bass drum track as well as the computed measure-wise similarity between both tracks are illustrated. Overall, we obtain a 143 dimensional feature space. The dimensions correspond to features that characterize the overall tempo (2 dim.), rhythmic properties of the single tracks **BD** (25), **SR** (25), **CH** (25), **BS** (25), and **GT** (25), as well as the similarity between the tracks **BS** & **BD** (4), **BS** & **SR** (4), **BS dim.** & **CH** (4), and **BS** & **GT** (4).

4.5. Regression Analysis

As mentioned before in Sec. 4.2, all 8 attributes are related to different instruments configurations. For each of the 8 regression models, we chose all features associated to the corresponding instrument(s). The number of features used for each model is given in the last column of Tab. 1 since no feature selection is performed in experiment **E3**. We compared the Stepwise Multiple Linear Regression (MLR) in experiment **E1** and the Partial-Least Squares Regression (PLS) in experiments **E2** and **E3** for modeling the attributes based on the annotated ground truth data and the extracted features for each segment.

The MLR performs a step-wise selection of features according to their statistical significance in a regression model for each attribute. If the features are assumed to be highly correlated, the PLS derives a smaller number of new predictor variables from a linear combination of the original input features. Here, the PLS takes the observed variability of the predictor variables (attribute values) into account. Since the PLS regression does not perform a feature selection, we investigate the effect of a preceding Principal Component Analysis (PCA) for feature space reduction in experiment **E2**. For more details on the regression methods, see [10], and for a recent application to musical features, see [7].

Regarding the small number of samples (45 segments), we performed a leave-one-out cross validation scenario to avoid overfitting. In each of the 45 folds, 44 segments were used to train the regression models and the averaged annotation of one segment were used to test the model. Therefore, we compute the root mean squared error (RMSE) between the observed and estimated attribute values to measure the predictive power of a regression model. When using PCA, we use all principal components with eigenvalues > 1. We use the Akaike information criterion to compute the optimal model order for the PLS regression.

5. RESULTS

As depicted in Tab. 1, the PLS regression without PCA achieved the lowest mean RMSE value and thus outperformed the other two configurations (MLR, PCA+PLS). One major drawback of the PLS regression model is that it depends on all input features which comes along with higher computational effort. The average model order of the PLS model with PCA is higher than for the MLR model. For a real-time application scenario, one would preferably use

Attribute	**Instrument(s)**	**Onset-classes**	**Average RMSE**			**Average model order**			**Average number of features**		
			E1	**E2**	**E3**	**E1**	**E2**	**E3**	**E1**	**E2**	**E3**
Interplay quality	All	**SR**, **BD**, **CH**, **BS**, **GT**	1.98	0.67	0.59	9	9	1	9	26	143
Groove quality	Drums, Bass	**SR**, **BD**, **CH**, **BS**	1.06	0.58	0.45	5	6	5	5	23	118
Pulse clarity	Bass	**BS**	1.15	0.84	0.37	2	6	4	2	11	43
Predictability of rhythm	Bass	**BS**	2.07	1.48	0.87	1	4	2	1	11	43
Pulse clarity	Drums	**SR**, **BD**, **CH**	1.40	1.02	0.78	6	3	3	6	18	89
Predictability of rhythm	Drums	**SR**, **BD**, **CH**	2.86	1.63	1.13	3	1	1	3	18	89
Pulse clarity	Guitar	**GT**	1.97	1.37	1.27	2	2	2	2	9	31
Predictability of rhythm	Guitar	**GT**	2.71	1.99	1.85	2	2	2	2	9	31
Mean values over all attributes			1.90	1.20	0.91	3.75	4.13	2.50	3.75	15.63	73.38

Table 1. Results of regression analysis for the experiments **E1** (Stepwise Multiple Linear Regression), **E2** (Partial Least-Squares Regression with preceding PCA), and **E3** (Partial Least-Squares Regression without PCA). See Sect. 4.3 and Sect. 4.5 for details. All values are averaged over 45 folds of the leave-one-out cross validation.

the MLR approach since it depends on a significantly lower number of features even though it leads to higher estimation errors.

6. CONCLUSIONS

In this paper, we presented a feature-based approach for modeling attributes to characterize musical ensemble recordings. The results indicate that it is generally possible to use the presented methods to estimate values of musical attributes that can be easily interpreted by musicologists without having detailed knowledge about the underlying algorithms. We achieved the lowest correct estimation rate for the guitar-related attributes. Another interesting finding is that values of *groove quality* (related to the bass and the drum) seems to be easier to estimate than those of the overall *interplay quality* of the complete trio.

7. ACKNOWLEDGEMENTS

The authors would like to thank the musicians and participants of the music annotation. This work has been partly supported by the German research project *GlobalMusic2One*[3] 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*[4], enabling transnational cooperation between Thuringian companies and their partners from other European regions.

8. REFERENCES

[1] G. Widmer and W. Goebl, "Computational models of expressive music performance: The state of the art," *Journal of New Music Research*, vol. 33, no. 3, pp. 203–216, 2004.

[2] H. Honing and W. B. De Haas, "Swing Once More: Relating Timing And Tempo In Expert Jazz Drumming," *Music Perception*, vol. 25, no. 5, pp. 471–476, 2008.

[3] Guy Madison, "Experiencing groove induced by music: consistency and phenomenology," *Music Perception*, vol. 24, pp. 201–207, 2006.

[4] M. Doffman, *Feeling the groove: shared time and its meanings for three jazz trios*, Ph.D. thesis, Mar. 2008.

[5] E. Stamatatos and G. Widmer, "Automatic identification of music performers with learning ensembles," *Artificial Intelligence*, vol. 165, pp. 37–56, 2005.

[6] O. Lartillot, T. Eerola, P. Toiviainen, and J. Fornari, "Multi-feature modeling of pulse clarity: design, validation and optimization," in *Proc. of the International Conference on Music Information Retrieval, Philadelphia, USA*, 2008.

[7] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR), Kobe, Japan*, 2009.

[8] C. Dittmar, K. Dressler, and K. Rosenbauer, "A toolbox for automatic transcription of polyphonic music," in *Proceedings of the Audio Mostly Conf., Ilmenau, Germany*, 2007.

[9] J. Abeßer, H. Lukashevich, P. Bräuer, and G. Schuller, "Bass playing style detection based on high-level features and pattern similarity," in *Proc. of the Int. Society of Music Information Retrieval (ISMIR), Utrecht, Netherlands*, 2010.

[10] I. S. Helland, "Partial least squares regression and statistical models," *Scandinavian Journal of Statistics*, vol. 17, no. 2, pp. 97–114, 1990.

[3] see http://www.globalmusic2one.net

[4] see http://www.songs2see.net