Audio content analysis for the prediction of musical performance properties

Konrad Krenzlin

Master's thesis

Audio content analysis for the prediction of musical performance properties

Masterarbeit
Studiengang Audiokommunikation und -technologie

Technische Universität Berlin
Fakultät I - Geisteswissenschaften
Fachgebiet Audiokommunikation

vorgelegt von: Konrad Krenzlin
Matrikelnummer: 309839

Erstgutachter: Prof. Dr. Stefan Weinzierl
Zweitgutachter: Dr. Steffen Lepa

eingereicht am: 21. Januar 2019

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt gegenüber der Fakultät I der Technischen Universität Berlin, dass die vorliegende, dieser Erklärung angefügte Arbeit selbstständig und nur unter Zuhilfenahme der im Literaturverzeichnis genannten Quellen und Hilfsmittel angefertigt wurde. Alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind kenntlich gemacht. Ich reiche die Arbeit erstmals als Prüfungsleistung ein. Ich versichere, dass diese Arbeit oder wesentliche Teile dieser Arbeit nicht bereits dem Leistungserwerb in einer anderen Lehrveranstaltung zugrunde lagen.

Mit meiner Unterschrift bestätige ich, dass ich über fachübliche Zitierregeln unterrichtet worden bin und verstanden habe. Die im betroffenen Fachgebiet üblichen Zitiervorschriften sind eingehalten worden. Eine Überprüfung der Arbeit auf Plagiate mithilfe elektronischer Hilfsmittel darf vorgenommen werden.

Berlin, den 21. Januar 2019

Konrad Krenzlin

Abstract

The main goal of this master's thesis was to generate regression models that can predict and explain perceived musical performance properties from audio content features. For this thesis 69 recordings, drawn from four musical fragments, and 9440 ratings of their performance properties were examined. Audio features measuring tempo, loudness and timbre were extracted and different regression methods compared. The resulting models were evaluated using Monte-carlo-cross-validation. Multiple mixed model analyses on the effects of composition and the raters were conducted.

Zusammenfassung

Das Ziel dieser Masterarbeit war es Regressionsmodelle zu entwickeln, die mittels technischer Signalmaße einer Aufnahmen Merkmale der musikalischen Interpretation vorhersagen und erklären können. Die extrahierten Signalmaße charakterisierten Tempo, Lautheit und Timbre. Dafür wurden 69 Aufnahmen von vier Stücken und 9440 Bewertungen ihrer musikalischen Aufführungsmerkmale betrachtet. Gleichzeitig wurden vier Regressionsmethoden verglichen. Mittels Monte-Carlo-Kreuzvalidierung wurden die resultierenden Modelle evaluiert. Anschließend wurde mit Hilfe von Mixed-Models der Einfluss der Werke bzw. der Bewerter untersucht.

Acknowlegdements

Thank you, Laura, for pushing me through this and having my back all the time.
This thesis would not have been possible without you.

I would like to thank Stefan Weinzierl and Steffen Lepa for their guidance, support and patience.

Last but not least, I would like to thank Heinz von Loesch for introducing me to this subject and starting this journey.

Contents

1. Introduction **3**

Theoretical Model . . . 3

Current Research . . . 3

2. Methods **5**

Musical stimuli . . . 5

Onset Annotation . . . 6

Listening Test . . . 6

Vocabulary . . . 6

Listening Tests . . . 7

Missing Values . . . 8

Intraclass Correlation Coefficient . . . 8

Feature Extraction . . . 11

Tempo and Timing . . . 11

Loudness and Dynamics . . . 11

Timbre . . . 12

Descriptive Statistics . . . 13

Standardisation and Functional Transformations . . . 13

Regression Models . . . 14

Intercept Model . . . 14

Linear Regression . . . 14

Stepwise Regression . . . 15

Regression Tree . . . 15

Random Forest Regression . . . 16

Evaluation . . . 16

Holdout set . . . 16

Model Selection . . . 16

Final Evaluation . . . 17

Interpretation . . . 17

Parsimonious models . . . 17

Mixed models 18

3. Results **19**

Comparing Regression Methods 19
Comparing Performance Attributes 21
Holdout Set 21
Cross-validation 22
Mixed models 23
Predictors 24

4. Discussion **31**

References **33**

List of Tables **37**

A. Recordings **39**

B. Questionnaire **43**

1. Introduction

Music, as conceived by a composer, is always perceived through the altering layer of a performer's rendition. Two performances based on the same musical score can have entirely different characteristics, yielding nuanced differences in the sonic qualities of a composition. While some perceived properties, such as tempo or loudness, correspond in the physical domain, more complex ones, such as articulation, expressiveness or phrasing, lack simple acoustical links. Still, there is a consensus among experts on the presence as well as the validity of these features.

This thesis mainly aims to uncover the extent to which it is possible to infer perceived musical properties of a performance from a mere audio signal. That is, are there sets of technical features that correlate with musical concepts of performance and how do they correlate?

Theoretical Model

If we consider the modified lens model developed by Juslin (1997), the performance of a composition can be viewed as a process of communication between the performer and the listener. First the performer's realisation of the score and his or her "expressive intention" is encoded using expressive acoustic cues, such as tempo, loudness or timbre. Then the listener uses these cues to form his or her own "judgements" and attributions of the performance.

Current Research

Since the advent of visual assessments of acoustic waves (Seashore, 1902), similarities and differences between musical performances have been studied based on these acoustic cues (Kendall & Carterette, 1990; Palmer, 1989; Repp, 1992a). Seashore (1967) also demonstrated that the variations are, to a great degree, consistent and repeatable.

The acoustic cues a listener uses to form his or her perception, i.e. the *functional validity*, have previously been studied for basic cues such as tempo and dynamics (Nakamura, 1987; Repp, 1994; Timmers, 2003). Considering the sometimes notable deviations between a listener's perception of performance properties and measured cues (Repp, 1992b), these basic cues alone might not sufficiently explain perception of a performance.

In order to expand on these ideas and identify more conclusive cues, Weinzierl & Maempel (2011) tried to predict 16 different performance properties, as rated by expert listeners. They selected the most important features of several hundreds of options, demonstrating, for example, the importance of note-by-note features.

2. Methods

This thesis builds on the dataset, used by Trebbin (2010), and Weinzierl & Maempel (2011) to cover three compositions by Beethoven, Mozart and Schumann and had previously been extended with a composition by Bach in 2017.

Musical stimuli

Drawn from the period between 1713 and 1849, four musical fragments were picked:

- Johann Sebastian Bach, Cello Suite No. 5 in C minor (BWV 1011), Sarabande, bars 1–20
- Ludwig van Beethoven, String Quartet No. 13 in B♭major (op. 130), 4th mov.: Alla danza tedesca – Allegro assai, bar 1–48
- Wolfgang Amadeus Mozart, Piano Sonata No. 12 in F major (K. 332), 1st mov.: Allegro, bars 41–94
- Robert Schumann, *Fünf Stücke im Volkston* for cello and piano (op. 102), 1st mov.: Mit Humor, bars 1–24

These fragments were selected, as they cover different instrumentations - cello, strings and piano - as well as genres - baroque, classical and romantic music. However, there is still common terminology used to describe the musical performances.

For each fragment, multiple recordings were selected: Mozart and Schumann 16, Beethoven 17 and Bach 20. In total 69 recordings from the period between 1939 and 2008 were used (c.f. appendix A).

The recordings were taken directly from CD maintaining a sample rate of $44,1kHz$ and a bit depth of $16bit$.

The durations of the recordings were, on average: 115s for Bach, 52s for Beethoven, 65s for Mozart and 30s for Schumann, allowing enough time for listeners to form an opinion (Thompson, Williamon, & Valentine, 2007) without becoming exhausted. In the case of the Bach composition, the recordings were edited by professionals who removed the repetition.

It should be recognised that the processes of recording and production themselves influences the performance properties (Lerch, 2011). Still, it can reasonably be said that, with the exclusion of loudness and its dynamics, some of the properties, such as tempo and rhythmisation, are solely related to performances, while others, such as timbre and dynamics, at least contribute greatly.

In order for this thesis to compensate for the influence of the production process on loudness and dynamics, all recordings were normalised for loudness according to Rec. ITU-R BS.1770-4.

Onset Annotation

For each musical stimulus, the timestamp of the onset of each note in the recording was manually annotated. Onsets were seen as the intended beginning of a note, which can differ from the actual realisation. For the initial dataset, onsets were obtained in a semi-automatic way. Onsets were detected by an audio-to-score-alignment algorithm described by Lerch (2008) and manually edited. Ornamentations, secondary voices and note values smaller than sixteenth notes were left out for simplification of the score. The additional recordings of the Bach fragment where annotated manually with the software Sonic Visualizer (Cannam, Landone, & Sandler, 2010).

All annotation files were checked for consistency, i.e. number and ordering of onsets, within each fragment.

Listening Test

Vocabulary

Previous to listening, a panel of expert listeners (N=10) that also rated the initial dataset was convened. The panel agreed on a common vocabulary to characterise the properties of the musical performances:

- Tone colour: tone colour of a single instrument or an ensemble independent of pitch or loudness
- Timbral bandwidth: variability of the tone colour caused by intonation
- Phrasing: width and strength of structuring of musical phrases
- Loudness: mean intended loudness
- Long-term dynamics: strength of differences in loudness between phrases
- Short-term dynamics: strength of differences in loudness within phrases
- Tempo: mean base tempo
- Agogic: bandwidth of tempo modulations within motifs and phrases
- Vibrato: periodic, minor pitch changes of a held note
- Rhythmisation: conciseness of depiction of the rhythm
- Articulation: strength of temporal separation between notes through shortened length, often also accentuated

Based on this vocabulary the following attributes and their poles were derived:

- Tone colour 1 (soft-hard)
- Tone colour 2 (dark-bright)

- Tone colour 3 (lean-full)
- Timbral bandwith (small-large)
- Phrasing width (narrow-wide)
- Phrasing strength (weak-strong)
- Loudness (gentle-loud)
- Long-term dynamics (weak-strong)
- Short-term dynamics (weak-strong)
- Tempo (slow-fast)
- Agogic (weak-strong)
- Rhythmisation (weak-concise)
- Articulation (legato-staccato)
- Articulation bandwidth (small-large)

Two attributes, describing general perception, were also added.

- Musical expression (weak-strong)
- Overall impression (dislike-like)

Vibrato was omitted, as it can only be rated for specific instrumentations. For example, a piano cannot produce a varying pitch.

Based on the attributes defined, a questionnaire was created for listeners to rate performances on five-point discrete scales. The values of the scales were chosen at random for prevention of ratings bias (c.f. appendix B)

Listening Tests

The initial dataset containing Beethoven, Mozart and Schumann was rated by a panel of 10 expert listeners (N=10, 2 female, 8 male) in 2010. The extension of Bach was rated by a second panel of five expert listeners (N=5, 5 male) in 2017.

Both listening tests were conducted at the studio of the *Audio Communication Group* at the Technical University of Berlin, which provided an ideal listening situation and met the requirements of Rec. ITU-R BS.1116-3.

All listeners were musical experts, being musicologists, producers, conductors or professional musicians.

In short discussions before the listening tests, participants reviewd the vocabulary and descriptions.

The participants were told which musical composition and instruments they would be listening to. They also were instructed to start filling out the questionnaire while listening.

After hearing one or more random stimuli, the participants agreed on an appropriate playback level. This level remained fixed for the rest of the listening session.

Each stimulus was pseudonymised by a group of characters. The participants were allowed to use pseudonyms throughout the test. They were instructed to note their pseudonyms and that of the stimulus in question in their questionnaires.

In a randomised order, all stimuli of a fragment were played back. After each stimulus, a small pause was given for participants to finish filling out their questionnaires.

The panel for the initial dataset was split into two groups. The second group listened to the stimuli in the reverse order to counteract possible sequential effects. As the sequences did not seem to affect the results, this was not done for the second panel.

The ratings for the initial listening test were only available as SPSS files. However, this proprietary format was converted into CSV files with the R library memisc.

The questionnaires for the second listening test were transcribed manually.

Ratings for *vibrato* were dropped, as they were only available for the Mozart fragment.

Missing Values

Overall, there were 9440 ratings, of which 46 (0,48%) were either undecided, i.e. in-between two items, undecipherable or not rated at all. These 46 ratings were considered missing.

Based on Rosenthal's guidelines (2017) Little's MCAR test (Little, 1988) was conducted. The test did not reject the null hypothesis of a *missing completely at random* pattern ($\chi^2(232) = 279.24, p = 0.018$)[1].

Given the small amount and the random pattern of the missing data, the values were imputed with the median of each attribute.

Intraclass Correlation Coefficient

In order for the listening test ratings to be applied to measure the attributes of the musical performances, the consistency between the experts' ratings needed to be quantified. A commonly used statistic is the *intraclass correlation coefficient* (ICC) (Shrout & Fleiss, 1979), which describes the degree of agreement between groups. Based on the classification of ICC (McGraw & Wong, 1996; Koo & Mae, 2016), a *two-way random effects models* testing for *consistency* of *multiple raters* was chosen, i.e. $ICC(C, k)$[2].

[1] The R library BaylorEdPsych was used.
[2] Since there was no implementation of $ICC(C, k)$ available, but the estimator is the same, $ICC(3, k)$ was used.

Table 2.1.: Intraclass correlation of panel ratings. A: initial panel, B: second panel

attribute	ICC(C,k) (A)	ICC(C,k) (B)
tone colour 1 (soft-hard)	0.83	0.71
tone colour 2 (dark-bright)	0.80	0.75
tone colour 3 (lean-full)	0.69	0.61
timbral bandwith (small-large)	0.76	0.66
phrasing width (narrow-wide)	0.31	0.52
phrasing strength (weak-strong)	0.62	0.56
loudness (gentle-loud)	0.78	0.56
long-term dynamics (weak-strong)	0.87	0.89
short-term dynamics (weak-strong)	0.70	0.59
tempo (slow-fast)	0.96	0.92
agogic (weak-strong)	0.90	0.79
rhythmisation (weak-concise)	0.70	0.62
articulation (legato-staccato)	0.83	0.75
articulation bandwidth (small-large)	0.77	0.77
musical expression (weak-strong)	0.77	0.79
overall impression (dislike-like)	0.76	0.86
all	0.79	0.75

Table 2.1 displays the ICC separately for both listening tests, as they were measured by different raters. Moreover, each item of the listening test was assessed individually[3]

Cicchetti (1994) described a coefficient less than 0.6 as poor or fair, a coefficient between 0.6 and 0.74 as good and one 0.74 above as excellent in terms of reliability. Based on these guidelines, the reliability of most of the attributes was good or excellent.

In contrast, the attributes *phrasing width* and *phrasing strength*, *loudness* and *short-term dynamics* only resulted in poor or fair reliability for at least one of the listening tests.

[3]With the exceptions of phrasing strength, agogic and rhythmisation, the results are the same as for Weinzierl & Maempel (2011).

The understanding of phrasing might not have been the same among the experts, which could have led to inconsistent ratings of width and strength.

The lack of reliability of the attribute *loudness* - mean loudness - can be explained by the fact that the stimuli had been previously normalised. Differences in perceived loudness, therefore, were less pronounced and subjective.

The fair reliability of *short-term dynamics*, the dynamics within a phrase, during the second listening test only, might have be due to difference among musical fragments.

As a result of the poor reliability in terms of measurement, the attributes *phrasing width*, *phrasing strength*, and *loudness* were dropped in further analyses. However *short-term dynamics* was kept as its reliability was almost good.

Multicollinearity

The attributes were suspected to be interdependent, which could have interfered with individual analyses of attributes.

Multicollinearity describes the circumstance in which a variable can, to a reasonable extent, be expressed by a linear combination of the other variables.

A commonly used statistic to quantify this phenomenon is the *variance inflation factor* (VIF).

Table 2.2.: variance inflation factors of performance attributes

attribute	VIF
tone colour (soft-hard)	1.32
tone colour (dark-bright)	1.22
tone colour (lean-full)	1.23
timbral bandwith (small-large)	1.57
long-term dynamics (weak-strong)	1.62
short-term dynamics (weak-strong)	1.33
tempo (slow-fast)	1.26
agogic (weak-strong)	1.39
rhythmisation (weak-concise)	1.34
articulation (legato-staccato)	1.49
articulation bandwidth (small-large)	1.77

musical expression (weak-strong)	2.07
overall impression (dislike-like)	1.42

Table 2.2 displays the VIF for each attribute.

As a general rule, values greater than 10 indicate high multicollinearity. In this study, the attributes did not exhibit problems of multicollinearity.

In conclusion, the attributes can be seen as independent and uncorrelated, and therefore, they can be analysed separately.

Feature Extraction

For representation of the performance attributes for each musical stimulus the following technical features characterising tempo and timing, loudness and dynamics and timbre were extracted.

Tempo and Timing

Tempo is usually measured in *beats per minute* (BPM).

$BPM = \frac{beats\ in\ quarters}{time\ in\ minutes}$

Onset times of two adjacent notes were used to compute inter-onset intervals (IOI) for each note. In combination with the distance of these two notes in the score, including rests, a local, micro tempo was calculated.

$BPM(i) = \frac{beat_{i+1} - beat_i}{onset_{i+1} - onset_i} * 60 \frac{sec}{min}$

The last notes had no following onsets. Therefore, no IOI and micro tempo could be calculated. Though mapping of each note to its local tempo, a tempo map was created. Since tempo changes are not perceived in absolute values, but rather in relative terms, another time series using the ratio of adjacent notes was derived.

$BPM_{ratio}(i) = \frac{BPM(i+1)}{BPM(i)}$

A value of 1.0 indicates no change in tempo, 1.1 indicates an increase of 10%, and 0.9 indicates a decrease of 10%.

Loudness and Dynamics

Each stimulus was loaded with *librosa* (McFee et al., 2015) and the original sample rate was maintained. In order to obtain a mono signal, the stereo signals were averaged.

Filtering according to Rec. ITU-RBS.1770-4 was applied, as this method has demonstrated (Soulodre, 2004) as having the highest correlation with perceptual loudness among existing loudness measures.

Based on the annotated onset times each stimulus was cut into slices corresponding to a single note.

ITU Loudness

Based on the formulas in Rec. ITU-RBS.1770-4 for a single channel signal, loudness was calculated for each note as follows:

$$L_i = -0.691 + 10\, log_{10} \left(\frac{1}{T_i} \int_0^{T_i} y_i^2 dt \right)$$

where y_i is the slice of the i-th note and T_i its length.

Decrease

In an attempt to measure the amount of legato play, the decrease in loudness from the maximum to the following minimum was measured.

$$decrease = L_{max} - L_{min}$$

where

$$i_{min} > i_{max}$$

It was hypothesised that a high value indicates a separation of the notes, i.e. staccato, and a low value indicates a compound, legato play.

Timbre

In contrast to tempo and loudness, timbre is a multidimensional property that incorporates both spectral and temporal patterns. This multidimensional aspect raises question of how many dimensions and with which features timbre can be measured.

Using multidimensional scaling Grey (1977) found three dimensions correlating with “spectral energy distribution”, the “synchronicity of higher harmonic transients”, along with the “amount of spectral fluctuation” and the “high-frequency energy in the initial attack segment”.

McAdams, Winsberg, Donnadieu, De Soete, & Krimphoff (1995) also found a three-dimensional model correlated with the logarithmic rise-time, the spectral centroid, and the degree of spectral variation.

Lakatos (2000) only identified the centroid and the rise time as principal dimensions of timbre.

Caclin, McAdams, Smith, & Winsberg (2005) also supported spectral centroid and attack time as the main dimensions, reporting spectral flux as a "less salient timbre parameter" with its influence depending on changes in other dimensions. Moreover, they added the "spectrum fine structure" as a fourth dimension.

While there is no absolute consistency across the findings of these studies, there seems to be consensus about at least two dimensions of timbre: the spectral centroid, measuring the brightness or sharpness, and the attack time, measuring the impulsiveness of the signal. Spectral flux can be seen as a third dimension, which measures the variability of the spectrum.

With *essentia* (Bogdanov et al., 2013), the spectral centroid, the logarithmic attack time and the spectral flux were measured for each note.

The spectral centroid is the weighted mean of the spectrum, indicating its "centre of mass". The logarithmic attack time is measured based on the log_10 of the attack time of the signal envelope, defined as the rise time between 20% and 95% of the maximum envelope value. The spectral flux is defined as the $L2$-norm (Euclidean distance) between two consecutive magnitude spectra.

Descriptive Statistics

The seven extracted features (BPM, BPM ratio, ITU loudness, decrease, spectral centroid, spectral flux, logarithmic attack time) create time series mapping features for every note. As the performance attributes describe overall characteristics, these time series need to be summarised. Most of the attributes are related to a property's mean, e.g. tempo, loudness, or a property's range, e.g. dynamics or timbral bandwidth. Therefore, the time series were described with the following statistics measuring central tendency and dispersion.

- mean
- median
- standard deviation
- minimum
- maximum
- interquartile range (IQR)

Standardisation and Functional Transformations

Participants in the listening test were instructed to rate the performances in relation to each other, independent of the musical fragment. Therefore, the technical features were standardised separately

within each composition. However, grouped standardisation decreased the resulting scores (c.f. 4) and therefore was omitted and the data left unaffected.

Given the perceptual characteristics, there did not need to be a linear relation to the technical features. In order to describe non-linear relations, two additional transformations were chosen: the logarithmic function, taking into account Weber's law, and the reversed quadratic function, representing an optimum curve.

The logarithmic transformation of the feature X_i is defined as follows:

$$X_i\_log = log_{10}(X_i + a + 1)$$

As the logarithm is only defined for values greater than zero, the constant offset $a + 1$, where a is the minimum value of X_i, was added to reach a minimum value of 1.

The negative square transformation of the feature X_i is defined as follows:

$$X_i\_negsquare = -(X_i)^2$$

After a combination of seven basic features, six descriptive statistics and three functional transformations (including the linear relation) 126 features resulted.

Regression Models

The first primary objective of this thesis was to develop an optimal regression model for each musical attribute by comparing the performance of different regression techniques. The following regression methods were implemented.

Intercept Model

As a baseline for comparison of the regression model performances, an intercept regression model predicting the mean of each attribute was used. Any other model that does not surpass its performance can be seen as insufficient, as it does not model the data better than a straight line.

Linear Regression

An ordinary least squares linear regression[4] is the most common form of regression, as it is easy to implement and interpretable via its regression coefficients.

[4] LinearRegression from the Python library sklearn was used.

In this study, the expected high multicollinearity of the features or independent variables could have violated the linear regression's assumptions, leading to high sensitivity to random errors and large variance. Additionally, overfitting is a common problem with a high number of independent variables compared to the number of observations.

Stepwise Regression

A stepwise regression can be seen as a technique that "wraps" a regression model with a feature selection method (Guyon & Elisseeff, 2003). Through recursive addition to or removal from a subset of features and building of a regression model, a given scoring function is optimised.

In this study, forward selection was used.[5] Starting with an empty feature set the feature that minimised the Akaike information criterion (AIC) was added at each step. When no such feature was found, the process was completed, leaving the final model and its subset of features.

It should be noted that this procedure is often criticised as it is prone to overfitting and can be biased as the same data is used for model building and selection.

Akaike information criterion (AIC)

The Akaike information criterion is a score used to evaluate the quality of different statistical models. It favours a trade-off between goodness of fit and simplicity by penalising higher numbers of features (predictors).

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

where k is the number of features and $\hat{L}$ the likelihood of the model.

A stepwise regression itself can be quite intensive to compute when there are many features, as many subsets have to be evaluated. A pure Python implementation, such as SequentialFeatureSelection (Raschka, 2018), the calculation of a single model using all features took about three minutes. With the use of stepAIC of the R library MASS and making the function callable from within the Python setup, the time of a single computation decreased to around 12 seconds.

Regression Tree

Regression trees are types of decision trees, models that build upon "if-then-else" rules. In the case of regression, the ending leaf nodes predict real values instead of a class label. Since the rules can become over-complex, they are prone to overfitting. There exist some strategies to prevent this, i.e. pruning, maximum depth or minimum node size. Finding the optimal decision tree is NP-complete (Hyafil &

[5] stepAIC from the R library MASS (Venables & Ripley, 2002) was used.

Rivest, 1976). Therefore, heuristic techniques, such as *Classification and Regression Trees* (CART), are used.

In this thesis, the DecisionTreeRegression of the Python library scikit-learn was used. No hyperparameter tuning was done, using the default parameters, resulting in full trees with maximised splits using all features.

Random Forest Regression

Random forest regression is an ensemble method, where instead of a single tree, multiple trees, i.e. the ensemble, are trained, and their predictions are averaged. Typical numbers of trees are between 10 and 100. The use of bootstrapped data and random subsets of features for each tree (Breiman, 2001) helps avoid overfitting and reduces the variance.

In this thesis, the RandomForestRegressor of the Python library scikit-learn was used. No hyperparameter tuning was done, and the default parameters were used, i.e. fully grown and unpruned trees. The number of trees was set to 100 instead of the default value of 10.

Evaluation

Given the relatively small sample size compared to the available features or independent variables, a thorough evaluation of the performance of the regression models was essential.

Holdout set

Using the same data to train and test a model's performance leads to highly biased estimates of the true prediction error on new, unseen data. Therefore 20% of the available data ($N = 590, N_{holdout} = 118$) was set aside and used for evaluation of the final models, i.e. the holdout method. Data were drawn stratified by stimuli, which ensured equal ratios for training and testing.

Model Selection

The performance and generalisation error of the different regression methods were estimated using Monte-Carlo-Cross-Validation (MCCV) (Xu & Liang, 2001, @Picard1984).

Based on the idea of the holdout method, in the MCCV the holdout set is repeatedly bootstrapped. For each training and holdout set the result of the scoring function is then calculated. Averaging these results yields more realistic estimates.

R^2 was used for the scoring function[6]. Contrary to its name, R^2 can be negative, denoting a model performing worse than one using only the mean.

Based on a comparison of the findings for each regression method and performance attribute, the optimal performing method, with regards to generalisation, for each performance attribute was selected.

In total 65000 models had to be trained and scored. The scikit-learn's parallelising capabilities were leveraged to reduce the total runtime of the MCCV to eight hours.

Final Evaluation

The results of the MCCV only provide a guideline for the selection of a model. An evaluation has to be done on new, unseen data.

Therefore, the regression models were trained on the complete training set and tested on the holdout set.

However, a score from a single holdout set can be overly pessimistic, underestimating the true performance of the model.

In order to prevent this, an additional 10-fold cross validation using the complete data, both the training and the holdout sets, was performed. This is a trade-off between the bias and variance of the estimator.

Interpretation

The second objective of this thesis is to explain the relations between the technical features and perception of musical performance properties.

Therefore, parsimonious and simple models were needed for interpretation of these relations.

Parsimonious models

For each performance attribute, a linear regression model that uses the features subset selected via the stepwise regression models was built

The data were standardised before the comparable regression coefficients were obtained.

Based on an examination of the features contributing the most, redundant linear combinations of similar features[7] were singled out manually. Since these combinations did contribute to the interpretation, they needed to be resolved by excluding one or more. Non-transformed was favoured over

[6] In the beginning, the mean squared error was used, as it does not assume linearity. However, no such discrimination was found during analysis, and the MSE was omitted due to simplicity.

[7] For example: 22.07 * ITU_std - 22.01 * ITU_IQR

transformed, mean over median, and standard deviation over interquartile range. As a result, 59 features were excluded.

Mixed models

For each performance attribute a linear regression model that uses the features subset selected via the stepwise regression models was built.

Then, a linear mixed effect analysis was performed through individual addition of random intercepts for the musical fragment and the rater. A likelihood ratio test was performed to test the significance of the mixed model.

The relative variance explained by the random intercepts was estimated by the ICC of the random-intercept-only model.

Based on Nakagawa & Schielzeth (2012), both the marginal R^2, describing the variance explained by fixed effects alone, and the conditional R^2, describing the variance explained by both fixed and random effects, were calculated.

The predictors were sorted by their beta coefficients to single out the most contributing predictors.

3. Results

Comparing Regression Methods

Table 3.1 displays the averaged R^2 results of the MCCV for each attribute and regression method.

Table 3.1.: mean of R^2 scores of MCCV (k=1000)

attribute	intercept	linear	stepwise	tree	random
tone colour 1 (soft-hard)	-0.01	-332375128.19	0.20	0.23	0.24
tone colour 2 (dark-bright)	-0.01	-542941512.25	0.12	0.16	0.17
tone colour 3 (lean-full)	-0.01	-476512716.34	0.13	0.12	0.13
timbral bandwith (small-large)	-0.01	-204618907.64	0.23	0.19	0.21
long-term dynamics (weak-strong)	-0.01	-566451572.43	0.33	0.34	0.35
short-term dynamics (weak-strong)	-0.01	-500470537.36	0.05	0.05	0.06
tempo (slow-fast)	-0.01	-150278333.69	0.64	0.66	0.67
agogic (weak-strong)	-0.01	-511538923.15	0.33	0.30	0.31
rhythmisation (weak-concise)	-0.01	-470969335.54	0.14	0.09	0.10
articulation (legato-staccato)	-0.01	-549652370.35	0.37	0.36	0.37
articulation bandwidth (small-large)	-0.01	-275840014.75	0.27	0.24	0.25
musical expression (weak-strong)	-0.01	-376557855.06	0.26	0.20	0.21
overall impression (dislike-like)	-0.01	-636093408.17	0.10	0.14	0.15

The intercept model R^2 scores were, as expected, near zero, as the mean value of each attribute is modeled.

For the linear regression, there were several huge outliers, due to extreme overfitting. These outliers result in substantial negative mean scores, resulting in models that are worse than the intercept models.

Both tree-based regression methods (tree and random forest) produced similar scores, with random forest being slightly better.

The scores of the stepwise regression were relatively the same as the tree-based methods. For timbral bandwidth, agogic, rhythmisation, articulation bandwidth and musical expression, the results were slightly better The results were worse for tone colour 1, tone colour 2 and overall impression.

Table 3.2.: median of R^2 scores of MCCV (k=1000)

attribute	intercept	linear	stepwise	tree	random
tone colour 1 (soft-hard)	-0.01	0.23	0.21	0.24	0.25
tone colour 2 (dark-bright)	-0.01	0.16	0.13	0.16	0.18
tone colour 3(lean-full)	-0.01	0.13	0.15	0.13	0.14
timbral bandwith (small-large)	-0.01	0.19	0.24	0.20	0.22
long-term dynamics (weak-strong)	-0.01	0.35	0.34	0.35	0.36
short-term dynamics (weak-strong)	-0.01	0.05	0.07	0.06	0.07
tempo (slow-fast)	-0.01	0.66	0.65	0.67	0.67
agogic (weak-strong)	-0.01	0.31	0.34	0.31	0.32
rhythmisation (weak-concise)	-0.01	0.10	0.15	0.10	0.11
articulation (legato-staccato)	-0.01	0.37	0.38	0.37	0.38
articulation bandwidth (small-large)	-0.01	0.24	0.28	0.25	0.25
musical expression (weak-strong)	-0.01	0.20	0.26	0.21	0.22
overall impression (dislike-like)	-0.01	0.14	0.11	0.15	0.16

Table 3.2 displays the median R^2 results of the MCCV scores. Since the outliers did not impact the median, the scores for the linear regression resembled the scores of the other methods. The results were consistent with the averaged scores.

Apart from the outliers of the linear regression, all regression methods produced approximately equal results.

Comparing Performance Attributes

The highest mean R^2 scores were for tempo, explaining 64% to 67% of its variance. The explained variance of long-term dynamics, agogic and articulation was between 30% and 37%.

With values below 10% short-term dynamics, rhythmisation and overall impression were not explained well.

The variance of the remaining attributes was explained by 20% to 30%.

For the attributes tempo, long-term dynamics, agogic and articulation the high scores corresponded with the ICC values 2.1. Conversely, this was not true for the lower scored attributes. Only the poor ICC can explain the worse performance for short-term dynamics. A consensus among the raters was necessary, but not sufficient, for a good prediction.

For tempo and loudness, there were direct correspondences with the features facilitating good prediction. Conversely, the overall impression was mainly subjective with no matching features.

Holdout Set

Table 3.3.: R^2 score of holdout set

attribute	intercept	linear	stepwise	tree	random
tone colour 1 (soft-hard)	-0.02	0.28	0.25	0.28	0.28
tone colour 2 (dark-bright)	-0.00	0.30	0.27	0.30	0.30
tone colour 3 (lean-full)	-0.00	0.01	0.12	0.01	0.01
timbral bandwith (small-large)	-0.00	0.45	0.47	0.45	0.45
long-term dynamics (weak-strong)	-0.00	0.40	0.42	0.40	0.40
short-term dynamics (weak-strong)	-0.00	0.11	0.12	0.11	0.11
tempo (slow-fast)	-0.01	0.68	0.68	0.68	0.68
agogic (weak-strong)	-0.00	0.34	0.30	0.34	0.35
rhythmisation (weak-concise)	-0.00	0.25	0.16	0.25	0.25
articulation (legato-staccato)	-0.01	0.42	0.44	0.42	0.42
articulation bandwidth (small-large)	-0.00	0.39	0.39	0.39	0.38
musical expression (weak-strong)	-0.00	0.30	0.29	0.30	0.31

overall impression (dislike-like)	-0.00	0.20	0.20	0.20	0.19

Table 3.3 displays the resulting R^2 scores of the predictions of the holdout set.

The intercept model R^2 scores were near zero, as expected.

Linear regression did not overfit in this case and produced scores in line with the regression tree.

Differences between the regression models were prominent for tone colour 3, where only stepwise regression explained about 12% of the variance. Additionally, the slightly higher mean scores for agogic and rhythmisation were now reversed.

With some exceptions, the results corresponded to the scores of the MCCV. There were only slightly higher values. Tempo was still the highest scoring attribute with around 68% explained variance.

The values for long-term dynamics and articulation were slightly higher, while the scores for agogic remained the same.

Prominent differences could be found in the high scores for timbral bandwidth, with around 46% explained variance, and the score of zero for tone colour 3.

Cross-validation

Table 3.4.: mean of R^2 scores of CV(k=10)

attribute	intercept	linear	stepwise	tree	random
tone colour 1 (soft-hard)	-0.02	0.26	0.21	0.26	0.27
tone colour 2 (dark-bright)	-0.04	0.22	0.20	0.23	0.25
tone colour 3 (lean-full)	-0.01	0.15	0.13	0.12	0.13
timbral bandwith (small-large)	-0.00	0.28	0.28	0.30	0.27
long-term dynamics (weak-strong)	-0.02	0.38	0.38	0.38	0.38
short-term dynamics (weak-strong)	-0.04	0.11	0.08	0.11	0.11
tempo (slow-fast)	-0.03	0.68	0.67	0.67	0.68
agogic (weak-strong)	-0.01	0.32	0.35	0.31	0.34
rhythmisation (weak-concise)	-0.02	0.15	0.22	0.17	0.14
articulation (legato-staccato)	-0.03	0.40	0.38	0.37	0.39

articulation bandwidth (small-large)	-0.02	0.31	0.27	0.30	0.29
musical expression (weak-strong)	-0.01	0.26	0.25	0.26	0.24
overall impression (dislike-like)	-0.01	0.16	0.20	0.16	0.17

Table 3.4 displays the mean R^2 scores of the 10-fold cross-validation, using both training and holdout sets.

The scores largely corresponded to the findings of the MCCV. While the differences in tone colour 2, timbral bandwidth, long-term dynamics and articulation bandwidth were prominent there is no clear tendency.

Mixed models

Table 3.5.: influence of random intercept: “fragment”

attribute	ICC_{iom}	R_c^2	R_m^2	$\chi^2(1)$	p
tone colour 1 (soft-hard)	0.14	0.30	0.30	0.00	1.00
tone colour 2 (dark-bright)	0.09	0.27	0.27	0.00	1.00
tone colour 3 (lean-full)	0.12	0.24	0.24	0.00	1.00
timbral bandwith (small-large)	0.28	0.38	0.38	0.00	1.00
long-term dynamics (weak-strong)	0.20	0.45	0.45	0.00	1.00
short-term dynamics (weak-strong)	0.11	0.18	0.18	0.00	1.00
tempo (slow-fast)	0.15	0.99	0.41	136.98	0.00
agogic (weak-strong)	0.16	0.40	0.40	0.00	1.00
rhythmisation (weak-concise)	0.19	0.26	0.26	0.00	1.00
articulation (legato-staccato)	0.37	0.47	0.47	0.00	1.00
articulation bandwidth (small-large)	0.32	0.42	0.42	0.00	1.00
musical expression (weak-strong)	0.21	0.38	0.38	0.00	1.00
overall impression (dislike-like)	0.12	0.27	0.27	0.00	1.00

Table 3.5 displays the results of the mixed models with a random intercept for the fragment, the musical composition.

A portion of variance was explained by the fragment (ICC_{iom}) for every attribute. For tempo only, this contribution led to a significant ($p < 0.01$) gain of 58%.

Table 3.6.: influence of random intercept: “subject”

attribute	ICC_{iom}	R_c^2	R_m^2	$\chi^2(1)$	p
tone colour 1 (soft-hard)	0.18	0.33	0.29	19.44	0.00
tone colour 2 (dark-bright)	0.10	0.28	0.27	1.07	0.30
tone colour 3 (lean-full)	0.17	0.28	0.24	16.78	0.00
timbral bandwith (small-large)	0.36	0.44	0.36	44.60	0.00
long-term dynamics (weak-strong)	0.26	0.51	0.44	58.02	0.00
short-term dynamics (weak-strong)	0.18	0.24	0.16	32.52	0.00
tempo (slow-fast)	0.15	0.64	0.64	1.01	0.31
agogic (weak-strong)	0.19	0.48	0.40	62.01	0.00
rhythmisation (weak-concise)	0.21	0.29	0.26	13.28	0.00
articulation (legato-staccato)	0.34	0.48	0.47	1.62	0.20
articulation bandwidth (small-large)	0.31	0.44	0.41	13.70	0.00
musical expression (weak-strong)	0.28	0.44	0.37	40.96	0.00
overall impression (dislike-like)	0.24	0.39	0.27	88.51	0.00

Table 3.6 display the results of the mixed models with a random intercept for the rater. Except for tone colour 2, tempo and articulation, there was a significant ($p < 0.01$) gain by considering the raters’ intercept.

Overall impression has the highest gain of about 12%, which makes sense as it is a highly subjective rating.

Moreover, timbral bandwidth, long-term and short-term dynamics, agogic and musical expression gained at least 6% to 8%.

Predictors

Table 3.7.: top predictors for “tone colour 1 (soft-hard)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.00	0.07	14.92	30.06	0.00
ITU_min	-0.38	0.09	589.71	-4.07	0.00
decrease_mean_log	-0.37	0.11	577.94	-3.51	0.00
spectral_flux_mean	-0.37	0.12	583.42	-3.17	0.00
ITU_mean	-0.36	0.17	589.75	-2.10	0.04

Table 3.8.: top predictors for “tone colour 2 (dark-bright)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.25	0.04	11.28	50.95	0.00
spectral_centroid_mean	1.26	0.46	573.91	2.77	0.01
spectral_centroid_median	-0.81	0.40	573.81	-2.00	0.05
decrease_mean_log	-0.59	0.16	574.09	-3.80	0.00
ITU_mean	-0.53	0.13	589.35	-4.05	0.00

Table 3.9.: top predictors for “tone colour 3 (lean-full)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.13	0.07	15.01	31.65	0.00
bpm_std	-0.69	0.14	578.48	-4.83	0.00
bpm_min	-0.40	0.13	576.76	-3.19	0.00
bpm_max	0.40	0.16	577.77	2.43	0.02
spectral_centroid_median	-0.32	0.05	577.49	-6.71	0.00

Table 3.10.: top predictors for “timbral bandwith (small-large)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.12	0.09	13.53	23.21	0.00
bpm_max_log	-0.84	0.22	582.57	-3.77	0.00
bpm_min	-0.67	0.15	574.44	-4.52	0.00
decrease_mean	0.60	0.17	579.62	3.55	0.00
ITU_max_log	-0.59	0.08	588.18	-7.33	0.00

Table 3.11.: top predictors for “long-term dynamics (weak-strong)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.23	0.09	13.90	24.66	0.00
ITU_std	1.32	0.18	574.59	7.25	0.00
bpm_mean	1.27	0.21	583.47	6.14	0.00
bpm_median_log	-0.95	0.22	578.80	-4.39	0.00
bpm_max_log	0.79	0.23	578.11	3.49	0.00

Table 3.12.: top predictors for “short-term dynamics (weak-strong)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.20	0.09	13.72	23.57	0.00
log_attack_time_max	0.33	0.07	535.99	4.64	0.00
bpm_ratio_IQR	0.30	0.07	579.64	4.07	0.00
decrease_mean_log	0.26	0.15	582.83	1.73	0.08
decrease_max	-0.22	0.08	584.98	-2.65	0.01

Table 3.13.: top predictors for “tempo (slow-fast)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
bpm_mean	2.48	0.26	577.05	9.66	0.00
(Intercept)	2.25	0.03	13.01	65.11	0.00
bpm_ratio_std	-1.99	0.27	576.07	-7.26	0.00
bpm_max	-1.84	0.54	576.05	-3.42	0.00
bpm_ratio_mean	1.65	0.22	576.05	7.45	0.00

Table 3.14.: top predictors for “agogic (weak-strong)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.20	0.04	590.00	56.27	0.00
bpm_median_log	-1.17	0.25	590.00	-4.73	0.00
bpm_max_log	0.72	0.24	590.00	2.98	0.00
bpm_IQR_log	0.59	0.18	590.00	3.32	0.00
spectral_centroid_mean	0.55	0.17	590.00	3.18	0.00

Table 3.15.: top predictors for “rhythmisation (weak-concise)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.21	0.06	15.21	37.74	0.00
bpm_mean	0.44	0.13	579.17	3.38	0.00
bpm_median_log	-0.38	0.22	577.08	-1.74	0.08
bpm_max_log	-0.31	0.20	580.60	-1.53	0.13
decrease_median	-0.24	0.07	579.61	-3.57	0.00

Table 3.16.: top predictors for “articulation (legato-staccato)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.11	0.05	9.66	46.24	0.00
bpm_min	0.72	0.11	573.10	6.29	0.00
bpm_median_log	-0.68	0.22	571.44	-3.06	0.00
decrease_max_log	0.53	0.11	571.51	5.00	0.00
bpm_std	0.46	0.16	570.91	2.83	0.00

Table 3.17.: top predictors for “articulation bandwidth (small-large)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.22	0.07	13.15	33.94	0.00
bpm_std	1.42	0.18	575.24	7.93	0.00
bpm_ratio_mean	-1.25	0.23	575.17	-5.52	0.00
bpm_IQR	-0.95	0.21	575.18	-4.52	0.00
bpm_max_log	-0.89	0.24	575.90	-3.67	0.00

Table 3.18.: top predictors for “musical expression (weak-strong)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	2.39	0.08	13.89	30.21	0.00
bpm_ratio_IQR	0.74	0.12	586.96	6.20	0.00
bpm_ratio_mean	-0.65	0.14	583.24	-4.67	0.00
bpm_min_log	-0.57	0.10	589.07	-5.78	0.00
bpm_std	0.45	0.14	574.97	3.12	0.00

Table 3.19.: top predictors for “overall impression (dislike-like)”

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1.89	0.12	15.38	15.22	0.00
spectral_centroid_median	-1.15	0.45	584.04	-2.56	0.01
spectral_centroid_mean	0.88	0.48	582.78	1.81	0.07
bpm_median_log	0.80	0.20	588.30	4.04	0.00
bpm_ratio_IQR	0.61	0.15	588.75	4.18	0.00

Tables 3.7 to 3.19 display the five predictors that contributed the most for each attribute. Except for tempo, the basis was the mixed model with a random intercept for raters.

There were contributing predictors that were associated with the respective attribute such as spectral centroid for tonal colour 2, the brightness, ITU loudness for long-term dynamics, mean BPM for tempo.

However, most of the predictors’ high beta values were due to suppressing effects of others and did not contribute to the interpretation.

4. Discussion

In trying to understand musical performance properties, an examination of how they relate to the acoustical domain is crucial.

This thesis has studied how signal-based audio features can predict attributes of performance properties. A total of 69 performances of four musical fragments was studied.

With the use of 126 features covering timing, loudness and timbre and different regression methods, models predicting expert ratings on 13 performance attributes were trained and evaluated.

Among the attributes only tempo was predicted well; more than 60% of its variance was explained. Second to tempo, only long-term dynamics, agogic and articulation followed with 30% to 37% explained variance.

A high ICC among the raters was necessary, but not sufficient for higher performance of an attribute.

In comparison to Weinzierl & Maempel (2011), the results were considerably lower. Only articulation, with 38% of its variance explained, performed better.

The results depended on if and how the data was standardized, especially for the stepwise regression. Table 4.1 displays the MCCV results when data was standardized within the fragments. The same effect occured, when centering the data within the groups. However, when standardizing the data across groups the results did not change.

Table 4.1.: mean of R^2 of MCCV(k=1000), standardised within fragment groups

attribute	intercept	linear	stepwise	tree	random
tone colour 1 (soft-hard)	-0.01	-15112646068092371206144	0.07	0.22	0.23
tone colour 2 (dark-bright)	-0.01	-1973666077439518834688	0.12	0.16	0.17
tone colour 3 (lean-full)	-0.01	-1531397065958860259328	0.04	0.12	0.13
timbral bandwith (small-large)	-0.01	-3013517571657203974144	-0.05	0.19	0.20
long-term dynamics (weak-strong)	-0.01	-56079934077941860794368	0.21	0.34	0.35
short-term dynamics (weak-strong)	-0.01	-4868881058580855259136	-0.03	0.05	0.06

tempo (slow-fast)	-0.01	-807931786231528030208	0.57	0.66	0.67
agogic weak-strong)	-0.01	-6666193206788588830720	0.18	0.30	0.31
rhythmisation (weak-concise)	-0.01	-10348324662249701507072	-0.04	0.09	0.10
articulation (legato-staccato)	-0.01	-9374840182395650441216	0.12	0.36	0.36
articulation bandwidth (small-large)	-0.01	-6220331702667967463424	-0.01	0.24	0.25
musical expression (weak-strong)	-0.01	-1245422671454085840896	0.03	0.20	0.21
overall impression (dislike-like)	-0.01	-2409428579903706497024	0.01	0.14	0.15

The analysis of the predictors did not further contribute to the interpretation of musical performances, Multicollinearity between the predictors seems to be the biggest issue.

In comparison to the different regression methods there were no substantial difference in performance, apart from anticipated overfitting problems. A regression method that incorporates the ordinal nature of the ratings, such as logistic ordinal regression, might have an advantage over assuming continuous variables.

Overall, while confirming the predictability of tempo and long-term dynamics, this thesis did not met its goal to gain further insights into more complex performance attributes. However, practical issues on features engineering and their standardisation were addressed and show the need for further research on how to handle the high multicollinearity among features.

References

Bogdanov, Dmitry; Nicolas Wack; Emilia Gómez; Sankalp Gulati; Perfecto Herrera; Oscar Mayor; Gerard Roma; Justin Salamon; José Zapata; and Xavier Serra (2013): "ESSENTIA: An Open-Source Library for Sound and Music Analysis." In: *Proceedings of the 21st Acm International Conference on Multimedia*, pp. 855-858. New York, NY, USA.

Breiman, Leo (2001): "Random Forests." In: *Machine Learning*, 45(1), pp. 5–32.

Caclin, Anne; Stephen McAdams; Bennett K. Smith; and Suzanne Winsberg (2005): "Acoustic Correlates of Timbre Space Dimensions: A Confirmatory Study Using Synthetic Tones." In: *The Journal of the Acoustical Society of America*, 118(1), pp. 471–482.

Cannam, C.; C. Landone; and M. Sandler (2010): "Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files." In : *Proceedings of the ACM Multimedia 2010 International Conference*, 1467–1468. Firenze, Italy.

Cicchetti, Domenic V. (1994): "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology." In: *Psychological Assessment*, 6(4), pp. 284–290.

Grey, John M. (1977): "Multidimensional Perceptual Scaling of Musical Timbres." In: *The Journal of the Acoustical Society of America*, 61(5), pp. 1270–1277.

Guyon, Isabelle; and André Elisseeff (2003): "An Introduction to Variable and Feature Selection." In: *Journal of Machine Learning Research*, 3(Mar), pp. 1157–1182.

ITU (2015): *Rec. ITU-R BS.1116-3: Methods for the subjective assessment of small impairments in audio systems.* Geneva: International Telecommunication Union

ITU (2015): *Rec. ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level.* Geneva: International Telecommunication Union

Hyafil, Laurent; and Ronald L. Rivest (1976): "Constructing Optimal Binary Decision Trees Is Np-Complete." In: *Information Processing Letters*, 5(1), pp. 15–17.

Juslin, Patrik N. (1997): "Emotional Communication in Music Performance: A Functionalist Perspective and Some Data." In: *Music Perception: An Interdisciplinary Journal*, 14(4), pp. 383–418.

Kendall, Roger A.; and Edward C. Carterette (1990): "The Communication of Musical Expression." In: *Music Perception: An Interdisciplinary Journal*, 8(2), pp. 129–163.

Koo, Terry K.; and Mae Y. Li (2016): "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." In: *Journal of Chiropractic Medicine*, 15 (2), pp. 155–163.

Lakatos, Stephen (2000): "A Common Perceptual Space for Harmonic and Percussive Timbres." In: *Perception & Psychophysics*, 62(7), pp. 1426–1439.

Lerch, Alexander (2008): *Software-Based Extraction of Objective Parameters from Music Performances*. Doctoral Thesis. Technische Universität Berlin, Fakultät 1, Fachgebiet Audiokommunikation, Berlin.

Lerch, Alexander (2011): "Musikaufnahmen als Datenquellen der Interpretationsanalyse." In: Heinz von Loesch (ed.) and Stefan Weinzierl (ed.) *Gemessene Interpretation - Computergestützte Aufführungsanalyse im Kreuzverhör der Disziplinen*, Mainz: Schott, pp. 157–171.

Little, Roderick J. A. (1988): "A Test of Missing Completely at Random for Multivariate Data with Missing Values." In: *Journal of the American Statistical Association*, 83(404), pp. 1198–1202.

McAdams, Stephen; Suzanne Winsberg; Sophie Donnadieu; Geert De Soete; and Jochen Krimphoff (1995): "Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specificities, and Latent Subject Classes." In: *Psychological Research*, 58(3), pp. 177–192.

McFee, Brian; Colin Raffel; Dawen Liang; Daniel P. W. Ellis; Matt McVicar; Eric Battenberg; and Oriol Nieto (2015): "Librosa: Audio and Music Signal Analysis in Python." In *Proceedings of the 14th Python in Science Conference*, Austin, Texas, USA, pp. 18–25.

McGraw, Kenneth O.; and S. P. Wong (1996): "Forming Inferences About Some Intraclass Correlation Coefficients." In: *Psychological Methods*, 1(1), pp. 30–46.

Nakagawa, Shinichi; and Holger Schielzeth n.d. "A General and Simple Method for Obtaining R2 from Generalized Linear Mixed-Effects Models." In: *Methods in Ecology and Evolution*, 4(2), pp. 133–142.

Nakamura, Toshie (1987): "The Communication of Dynamics Between Musicians and Listeners Through Musical Performance." In: *Perception & Psychophysics*, 41 (6), pp. 525–533.

Palmer, C. (1989): "Mapping Musical Thought to Musical Performance." In: *Journal of Experimental Psychology. Human Perception and Performance*, 15(2), pp. 331–346.

Picard, Richard R.; and R. Dennis Cook (1984): "Cross-Validation of Regression Models." In: *Journal of the American Statistical Association*, 79(387), pp. 575–583.

Raschka, Sebastian (2018): "MLxtend: Providing Machine Learning and Data Science Utilities and Extensions to Python's Scientific Computing Stack." In: *The Journal of Open Source Software*, 3(24), 638.

Repp, Bruno H. (1992a): "Probing the Cognitive Representation of Musical Time: Structural Constraints on the Perception of Timing Perturbations." In: *Cognition*, 44(3), pp. 241–81.

Repp, Bruno H. (1992b): "Diversity and Commonality in Music Performance: An Analysis of Timing Microstructure in Schumann's "'Träumerei"'." In: *The Journal of the Acoustical Society of America*, 92(5),

pp. 2546–2468.

Repp, Bruno H. (1994): “On Determining the Basic Tempo of an Expressive Music Performance.” In: *Psychology of Music*, 22(2), pp. 157–167.

Rosenthal, Sonny (2017): “Data Imputation.” In: *The International Encyclopedia of Communication Research Methods*, American Cancer Society, pp.1–12.

Seashore, Carl Emil (1902): “A Voice Tonoscope.” In: *University of Iowa Studies in Psychology*, 3, pp. 18–28.

Seashore, Carl Emil (1967): *Psychology of Music*. New York: Dover.

Shrout, P. E.; and J. L. Fleiss (1979): “Intraclass Correlations: Uses in Assessing Rater Reliability.” In: *Psychological Bulletin*, 86 (2), pp. 420–428.

Soulodre, Gilbert A. (2004): “Evaluation of Objective Loudness Meters.” In: *Journal of the Canadian Acoustical Association*, 32(4), pp. 152–153,

Thompson, Sam; Aaron Williamon; and Elizabeth Valentine (2007): “Time-Dependent Characteristics of Performance Evaluation.” In: *Music Perception: An Interdisciplinary Journal*, 25 (1), pp. 13–29.

Timmers, Renee (2003): *Predicting the Subjective Similarity Between Expressive Performances of Music from Objective Measurements of Tempo and Dynamics*. Technical Report OFAI-TR-2003-25. Österreichisches Forschungsinstitut für Artificial Intelligence, Wien.

Trebbin, Wilko (2010): “Zusammenhangsanalyse von perzeptiven Attributen und Audio-Features zur Beschreibung von musikalischen Interpretationen.” Magisterarbeit, Technische Universität Berlin, Fakultät 1, Fachgebiet Audiokommunikation, Berlin.

Venables, W. N.; and B. D. Ripley (2002): *Modern Applied Statistics with S*. 4th ed., New York: Springer.

Weinzierl, Stefan; and Hans-Joachim Maempel (2011): “Zur Erklärbarkeit der Qualitäten musikalischer Interpretationen durch akustische Signalmaße.” In: Heinz von Loesch (ed.) and Stefan Weinzierl (ed.) *Gemessene Interpretation - Computergestützte Aufführungsanalyse im Kreuzverhör der Disziplinen*, Mainz: Schott, pp. 213–236.

Xu, Qingsong; and Yi-Zeng Liang (2001): “Monte Carlo Cross Validation.” In: *Chemometrics and Intelligent Laboratory Systems*, 56(April), pp. 1–11.

List of Tables

2.1. Intraclass correlation of panel ratings. A: initial panel, B: second panel 9
2.2. variance inflation factors of performance attributes 10

3.1. mean of R^2 scores of MCCV (k=1000) 19
3.2. median of R^2 scores of MCCV (k=1000) 20
3.3. R^2 score of holdout set 21
3.4. mean of R^2 scores of CV(k=10) 22
3.5. influence of random intercept: “fragment” 23
3.6. influence of random intercept: “subject” 24
3.7. top predictors for “tone colour 1 (soft-hard)” 25
3.8. top predictors for “tone colour 2 (dark-bright)” 25
3.9. top predictors for “tone colour 3 (lean-full)” 25
3.10. top predictors for “timbral bandwith (small-large)” 26
3.11. top predictors for “long-term dynamics (weak-strong)” 26
3.12. top predictors for “short-term dynamics (weak-strong)” 26
3.13. top predictors for “tempo (slow-fast)” 27
3.14. top predictors for “agogic (weak-strong)” 27
3.15. top predictors for “rhythmisation (weak-concise)” 27
3.16. top predictors for “articulation (legato-staccato)” 28
3.17. top predictors for “articulation bandwidth (small-large)” 28
3.18. top predictors for “musical expression (weak-strong)” 28
3.19. top predictors for “overall impression (dislike-like)” 29

4.1. mean of R^2 of MCCV(k=1000), standardised within fragment groups 31

A.1. Bach 39
A.2. Beethoven 40
A.3. Mozart 40
A.4. Schumann 41

A. Recordings

Table A.1.: Bach

year	performer
1957	János Starker
1991	Mstislaw Rostropowitsch
2007	Jean-Guihen Queyras
1939	Ludwig Hoelscher
1965	Ludwig Hoelscher
1979	Anner Bylsma
1964	Enrico Mainardi
2005	Truls Mørk
1985	Mischa Maisky
1964	Maurice Gendron
1998	Pieter Wispelwey
1957	Gaspar Cassadó
1982	Yo-Yo Ma
1984	Heinrich Schiff
1982	Paul Tortelier
1982	Lynn Harrell
2000	Daniel Müller-Schott
1960	Pierre Fournier
2003	Alexander Kniazev
1971	Daniil Shafran

Table A.2.: Beethoven

year	performer	label
1970	Juilliard String Quartet	Sony
1941	Busch Quartett	Sony
1990	Lindsay String Quartet	ASV
1962	Amadeus Quartett	DGG
1994	Emerson String Quartet	DGG
1989	Alban Berg Quartett	EMI
1987	Guarneri Quartet	Decca
1957	Hollywood String Quartet	Testament
1985	Melos Quartet	DGG
1999	Petersen Quartet	Capriccio
1973	Végh Quartet	Valois
1972	LaSalle String Quartet	DGG
1969	Quartetto Italiano	Philips
1952	Végh Quartet	Music & Arts
1982	Smetana Quartet	Denon
1971	Yale Quartet	Brilliant
1990	Tokyo String Quartet	RCA

Table A.3.: Mozart

year	performer	label
1981	András Schiff	DECCA
1985	Daniel Barenboim	EMI Classics
1972	Glenn Gould	Sony Classics
1946	Vladimir Horowitz	Classica D'Oro
1967	Lili Kraus	Sony Classics

1983	Mitsuko Uchida	Phi
1984	Christian Zacharias	EMI Classics
1988	Mieczyslaw Horszowski	Elektra Nonesuch
2005	Mikhail Pletnev	DG
1953	Walter Gieseking	EMI Classics
1954	Lili Kraus	Music & Arts
2000	Alfred Bendel	Phi
2000	Lars Vogt	EMI Classics
1993	Maria Joao Pires	DG
1982	Friedrich Gulda	DG
1974	Cecile Ousset	Berlin Classics

Table A.4.: Schumann

year	performer	label
1987	Raphael Wallfisch, Peter Wallfisch	Chandos
1993	Jan Vogler, Bruno Canino	Berlin Classics
1997	Steven Isserlis, Christoph Eschenbach	RCA Red Seal
1995	Anner Bylsma, Lamber Orkis	Sony Classics
1988	Klaus Storck, Yasuko Matsuda	Colosseum
1993	Maria Kliegel, Kristin Merscher	Naxos
1999	Mischa Maisky, Martha Argerich	DG
1967	Pierre Fournier, Jean Fonda	DG
1961	Mstislav Rostropowitsch, Benjamin Britten	Decca
1995	Truls Mørk, Leif Ove Ansdnes	Simax
1988	Yo-Yo Ma, Emanuel Ax	SK
1952	Pablo Casals, Leopold Mannes	Sony Classics
2008	Jean-Guihen Queyras, Eric Le Sage	Alpha

1978	Friedrich Jürgen Sellheim, Eckhart Sellheim	Sony Classics
1978	Andre Navarra, Annie d’Arco	Calliope
2004	Daniel Müller-Schott, Robert Kulek	Orfeo

B. Questionnaire

Hörversuch Interpretationsanalyse

Im Folgenden präsentieren wir Ihnen jeweils einen Ausschnitt aus einem Musikstück. Uns interessiert Ihre Einschätzung der verschiedenen Interpretationen des Stückes. Bitte füllen Sie im Verlauf der einzelnen Darbietungen jeweils einen Fragebogen vollständig aus.

Zur Erläuterung der im semantischen Differential genannten Merkmale:

Klangfarbe:	Klangcharakter eines einzelnen Instruments (bei Soloaufnahmen) oder eines Instrumentalensembles unabhängig von Tonhöhe und Lautstärke. ► weich — hart ► hell — dunkel ► schlank — voll
Klangfarbliche Bandbreite:	Veränderlichkeit der durch die Tongebung der Musiker bedingten Klangfarbe. ► groß — klein
Phrasierung:	Dauer und Prägnanz der durch den Spieler vorgenommenen Gliederung des musikalischen Verlaufs in musikalische Abschnitte (Phrasen). ► kleinteilig — weiträumig ► stark — schwach
Lautstärke:	Die mittlere mutmaßlich intendierte Lautstärke der Interpreten ► leise — laut
Dynamische Bandbreite:	Größe der Lautstärkeunterschiede zwischen verschiedenen Phrasen. ► gering — hoch
Binnendynamik:	Größe der Lautstärkeunterschiede innerhalb einzelner Phrasen: ► gering — hoch
Tempo:	Mittleres Grundtempo der Interpretation ► langsam — schnell
Agogik:	Bandbreite der Tempomodulation um ein bestimmtes Tempo (innerhalb von Motiven und Phrasen) ► wenig — viel
Vibrato:	Periodisch wiederkehrende, geringfügige Veränderung der Frequenz eines gehaltenen Tones. ► wenig — viel
Rhythmisierung:	Deutlichkeit der Darstellung eines durch den Notentext vorgegebenen Rhythmus. ► prägnant — unprägnant
Artikulation:	Grad der zeitlichen Trennung von Noten durch Verkürzung ihrer Spieldauer – häufig unterstützt durch stärkere Betonung. ► abgesetzt — gebunden
Artikulatorische Bandbreite:	Unterschiedlichkeit in der Artikulation einzelner Töne über den gesamten musikalischen Verlauf. ► groß — klein

ID:

Klangfarbe	weich ○ ○ ○ ○ ○ hart	
	hell ○ ○ ○ ○ ○ dunkel	
	schlank ○ ○ ○ ○ ○ voll	
Klangfarbliche Bandbreite	groß ○ ○ ○ ○ ○ klein	
Phrasierung	kleinteilig ○ ○ ○ ○ ○ weiträumig	
Phrasierung	stark ○ ○ ○ ○ ○ schwach	
Lautstärke	leise ○ ○ ○ ○ ○ laut	
dynamische Bandbreite	gering ○ ○ ○ ○ ○ hoch	
Binnendynamik	gering ○ ○ ○ ○ ○ hoch	
Tempo	langsam ○ ○ ○ ○ ○ schnell	
Agogik	wenig ○ ○ ○ ○ ○ viel	
Vibrato	viel ○ ○ ○ ○ ○ wenig	
Rhythmisierung	prägnant ○ ○ ○ ○ ○ unprägnant	
Artikulation	abgesetzt ○ ○ ○ ○ ○ gebunden	
Artikulatorische Bandbreite	groß ○ ○ ○ ○ ○ klein	
Musikalischer Ausdruck	schwach ○ ○ ○ ○ ○ stark	
Gesamteindruck	gefällt mir nicht ○ ○ ○ ○ ○ gefällt mir	