Stefan Weinzierl · Hans-Joachim Maempel

Zur Erklärbarkeit der Qualitäten musikalischer Interpretationen durch akustische Signalmaße

Versuche, spezifische Charakteristika von musikalischen Interpretationen aus dem akustischen Signal der Aufführung herauszulesen, haben eine lange Tradition. Sie beginnt spätestens mit den Arbeiten der Gruppe um den Musikpsychologen Carl E. Seashore, die bereits um 1900 auf der Grundlage von neuen Apparaturen zur Sichtbarmachung von akustischen Schwingungen[1] Tonhöhenverläufe und das Vibrato von Gesangsstimmen analysierte.[2] Ende der 1970er Jahre erlebt die Interpretationsforschung einen weiteren Aufschwung durch digitale Aufzeichnungstechnologien, mit denen Schallsignale leichter visualisiert und analysiert werden konnten. Im Gegensatz zu Messungen der motorischen Abläufe der Klangerzeugung, wie sie Seashores Piano Camera,[3] die in den 1980er Jahren entwickelten Klaviere mit optischen MIDI-Interfaces[4] oder moderne Sensortechnologien am Instrument des Spielers liefern, verspricht die Analyse der akustischen Signale einen engeren Bezug zum Höreindruck der musikalischen Aufführung.

Die digitale Aufzeichnung von Musik wurde im Rahmen der Interpretationsforschung zunächst überwiegend für eine bessere Visualisierung eingesetzt, etwa um Notenanfänge und die daraus resultierende zeitliche Struktur der Aufführung zu identifizieren.[5] Mit der zunehmenden Leistungsfähigkeit digitaler Signalverarbeitung wird diese manuelle Annotation allerdings zunehmend ergänzt durch die algorithmische Berechnung akustischer Features, in denen bis zu einem gewissen Grad bereits Modelle für die Wahrnehmung klanglicher Eigenschaften und elementarer musikalischer Ereignisse implementiert sind. Dazu gehören Algorithmen für die Erkennung von Notenanfängen (Onset detection), Prädiktoren für elementare

1 Carl E. Seashore, ›A Voice Tonoscope‹, in: *Universtity of Iowa Studies in Psychology* 3 (1902), S. 18–28.

2 Max Schoen, ›Pitch and Vibrato in Artistic Singing. An Experimental Study‹, in: *Musical Quarterly* XII/2 (1926), S. 275–290.

3 Joseph Tiffin und Carl E. Seashore, ›The Iowa Piano Camera‹, in: *Science* 72/1858 (1930), S. 146–147.

4 Caroline Palmer und Judith C. Brown, ›Investigations in the Amplitude of Sounded Piano Tones‹, in: *Journal of the Acoustical Society of America* 90/1 (1991), S. 60–66.

5 Bruno H. Repp, ›Diversity and Commonality in Music Performance: An Analysis of Timing Microstructure in Schumann's »Träumerei«‹, in: *Journal of the Acoustical Society of America* 92/5 (1992), S. 2546–2568.

sensorische Qualitäten wie die Lautstärke und die Klangfarbe von akustischen Signalen, sowie Algorithmen zur Erkennung der Tonhöhe und ihres Verlaufs. Zur Zeit werden diese Prädiktoren überwiegend für Anwendungen im Bereich des Music Information Retrieval, etwa zur Beschreibung von Musik in großen Datenbanken, zur Identifikation von Melodieanfängen oder zur Klassifikation musikalischer Genres verwendet. Allerdings bieten diese Features ein umfangreiches Repertoire signalbasierter Deskriptoren, das in jüngster Zeit auch zur Beschreibung von Merkmalen musikalischer Interpretationen eingesetzt wurde.

Ausgehend von einer Identifikation musikalischer Einzelereignisse und der Beschreibung von Momentanwerten für den Zeitverlauf von Lautstärke, Klangfarbe und Tonhöhe wurde hierbei untersucht, inwieweit integrative Konzepte des musikalischen Ausdrucks wie ein ›mittleres Tempo‹ (mean tempo)[6] bzw. ein ›Grundtempo‹ (main tempo)[7] aus dem zeitlichen Verlauf oder der statistischen Verteilung der Einzelwerte abgeleitet werden können. Erstaunlich wenige Studien haben sich bisher allerdings mit der Frage befasst, inwieweit nicht nur einzelne Parameter wie Tempoverläufe oder dynamische Gestaltungsmerkmale, sondern die Gesamtheit dieser Merkmale aus der akustischen Information herausgelesen werden kann – inwieweit also das Spezifische einer Interpretation im Gegensatz zu anderen Realisierungen des gleichen Notentexts durch die Dimensionen, die üblicherweise zur Beschreibung von akustischen Signalen verwendet werden (Zeitverlauf, Intensität, Spektrum), erfasst werden kann.

Vor diesem Hintergrund hat etwa Renee Timmers in einer 2005 veröffentlichten Studie untersucht, inwieweit die von Hörern bewertete Ähnlichkeit von Interpretationen desselben Werks durch eine Analyse von Tempo und Lautheit erklärt werden kann.[8] Präsentiert wurden relativ kurze, vier- bis zehntaktige Ausschnitte (Chopin, Mozart), die von Hörern jeweils paarweise im Hinblick auf ihre Ähnlichkeit bewertet wurden. Im Ergebnis konnten etwa 20 bis 30 Prozent der Varianz der empfundenen Unähnlichkeit mit einem Regressionsmodell durch eine Reihe von elementaren, auf Tempo und Lautheit bezogenen Parametern erklärt werden.

In einer an der Technischen Universität Berlin durchgeführten Untersuchung wurde versucht, einen Schritt weiterzugehen; zum einen, indem nicht nur Ähnlichkeiten zwischen, sondern differenzierte Merkmale von Interpretationen erhoben werden sollten, zum anderen, indem nicht nur zwei relativ elementare Tempo- und Lautheitsmaße, sondern vielfältige, auf dem akustischen Signal basierende Merkmale extrahiert wurden. Hierfür wurde eine Softwarebibliothek für den Einsatz signalbasierter Merkmale in der Interpretationsforschung entwickelt, die im Beitrag von

6 Bruno H. Repp, ›On Determining the Basic Tempo of an Expressive Music Performance‹, in: *Psychology of Music* 22/2 (1994), S. 157–167.

7 Alf Gabrielsson, ›The Performance of Music‹, in: Diana Deutsch (Hrsg.), *The Psychology of Music*, San Diego ²1999, S. 501–602.

8 Renee Timmers, ›Predicting the Similarity between Expressive Performances of Music from Measurements of Tempo and Dynamics‹, in: *Journal of the Acoustical Society of America* 117/1 (2005), S. 391–399.

Alexander Lerch in diesem Band vorgestellt wird.[9] Ziel war es also zu überprüfen, inwieweit die akustischen Merkmale entweder einzeln oder in Kombination geeignet sein können, die Höreindrücke zu erklären.

1. Musikalisches Material

In einem ersten Schritt wurden Expertenurteile zu verschiedenen Interpretationen des gleichen Werks erhoben. Die Hörbeispiele sollten einerseits lang genug sein, dass auch Merkmale wie Phrasierung, Agogik oder Dynamik beurteilt werden können, die sich auf einen größeren musikalischen Zusammenhang beziehen und für die Bewertung von Interpretationen zweifellos bedeutsam sind, andererseits sollte eine angemessene Gesamthörversuchsdauer eingehalten werden. Da gezeigt wurde, dass Experten ein Urteil über die Qualität einer Interpretation typischerweise innerhalb von etwa 20 s bilden und nach einer Minute kaum noch Änderungen daran vornehmen,[10] wurden 30 bis 60 s lange Ausschnitte aus folgenden Werken gewählt:

- Wolfgang Amadeus Mozart, Klaviersonate F-Dur KV 332, 1. Satz, T. 41–94,
- Ludwig van Beethoven, Streichquartett Nr. 13 B-Dur op. 130, 4. Satz, T. 1–48,
- Robert Schumann, *Fünf Stücke im Volkston* für Violoncello und Klavier op. 102, 1. Satz, T. 1–24.

Mit der Auswahl von Werken aus dem klassisch-romantischen Repertoire sollte gewährleistet werden, dass eine einheitliche Terminologie im Hinblick auf Vortrags- und Gestaltungsmerkmale verwendet werden kann. Gleichzeitig sollte durch die Berücksichtigung verschiedener Werke aus einem Zeitraum von 1783 (Mozart) bis 1849 (Schumann) eine gewisse Bandbreite im Hinblick auf Stilistik, Besetzung und musikalische Charaktere abgedeckt werden. In den ausgewählten Passagen waren keine ausgeprägten formalen Zäsuren in Form von neuen Tempi oder musikalischen Charakterbezeichnungen enthalten, die etwa eine eindeutige Einschätzung des mittleren Tempos nicht zulassen würden.

Für jedes Werk wurden 16 (Beethoven: 17) Einspielungen ausgewählt, die in der Zeit zwischen 1947 und 2007 entstanden sind (siehe Anhang). Bei den untersuchten ›Interpretationen‹ handelt es sich in Wahrheit natürlich um ›Aufnahmen von Interpretationen‹, die einen mehr oder weniger komplexen und für den Hörer ebenso wie für den Forscher im Nachhinein weitgehend intransparenten Prozess von Audiobearbeitungen durchlaufen haben.[11] Dies schränkt die Gültigkeit der Beurteilungen nicht ein, solange man anerkennt, dass mediale Interpretationen Ergebnis eines mehrstufigen Prozesses sind, wobei einige Merkmale wie Tempo und Agogik ganz

[9] ›Software-gestützte Merkmalsextraktion für die musikalische Aufführungsanalyse‹, S. 205 bis 212.

[10] Sam Thompson, Aaron Williamon und Elizabeth Valentine, ›Time-Dependent Characteristics of Performance Evaluation‹, in: *Music Perception* 25/1 (2007), S. 13–29.

[11] Siehe dazu den Beitrag von Hans-Joachim Maempel in diesem Band, S. 157–171.

überwiegend durch den musikalischen Interpreten festgelegt werden, während andere wie Klangfarbe, Lautstärke und Dynamik erheblich durch Eingriffe des Produzenten beeinflusst sind. Aus diesem Grund wurde mit einer Ausnahme auf Aufnahmen der Schellack-Ära verzichtet, wo tonträgerspezifische Artefakte (Knackser, Rauschen) einerseits zu fehlerhafter Onset-Erkennung führen können und andererseits einen erheblich stärkeren Einfluss auf Beurteilungen von Dynamik und Klangfarbe ausüben als die Eigenschaften späterer Aufzeichnungsmedien wie Magnetband oder digitale Tonträger.

Nur ein Aspekt des Produktionsprozesses wurde nachträglich kompensiert. Da Unterschiede in der Aussteuerung auf dem Wiedergabemedium einen starken Einfluss nicht nur auf die Bewertung dynamischer Eigenschaften, sondern auf die Gesamtheit aller Interpretationsparameter hätten, wurden alle Stimuli nach ITU-R BS.1770 in der Lautheit angeglichen. Im Ergebnis kann hierbei allerdings nicht nur die technische Aussteuerung, sondern auch eine unterschiedliche musikalische Auslegung der mittleren Dynamik kompensiert worden sein. Insofern können für Einschätzungen der ›Lautstärke‹ (s. u.) durchweg ähnliche Werte erwartet werden.

2. Hörversuch

Die Beurteilung der insgesamt 49 Einspielungen wurde einer Expertengruppe von zehn Personen anvertraut, darunter Dirigenten, Tonmeister, professionelle Musiker und Musikwissenschaftler – Personen also, die im Rahmen ihrer beruflichen Tätigkeit ständig Beurteilungen von musikalischen Interpretationen vornehmen. In einer moderierten Panel-Diskussion wurde zunächst ein konzeptspezifisches, semantisches Differential von im Ergebnis 16 bipolaren Attributen erarbeitet, wie sie typischerweise für die Beschreibung unterschiedlicher Interpretationen von Werken des klassisch-romantischen Repertoires verwendet werden (Tabelle 1).

Das Differential enthält sowohl Merkmale, die die mittlere Tendenz von Qualitäten über die Spieldauer beschreiben, etwa Tempo, Lautstärke, Klangfarbe oder Artikulation, als auch solche, die deren Veränderlichkeit erfassen, etwa klangfarbliche Bandbreite, Dynamik, Agogik oder artikulatorische Bandbreite. Einige Begriffe wurden zunächst nicht von allen Experten gleichbedeutend verwendet. Zum Teil wurde erst nach einer längeren Diskussion und gestützt durch Aufnahmen, wie sie im Anschluss auch im Hörversuch verwendet wurden, ein Konsens erreicht und durch Arbeitsdefinitionen festgehalten.

Im anschließenden Hörversuch sollten die Expertenhörer nur Unterschiede zwischen Interpretationsmerkmalen innerhalb desselben Werks beurteilen, also z. B. nicht das per Spielanweisung gegebene absolute Tempo, sondern das relative Tempo im Vergleich der präsentierten Interpretationen. Um dies zu ermöglichen, wurden jeweils einige Extrembeispiele im Hinblick auf Tempo, Agogik und dynamische Gestaltung im Voraus angespielt. Außerdem wurde das Panel in zwei Gruppen geteilt, die die Einspielungen jeweils in umgekehrter Reihenfolge hörten, um Reihenfolgen-

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

Tabelle 1: Semantisches Differential zur Bewertung der vorgelegten 49 Interpretationen von drei Werken des klassisch-romantischen Repertoires, als Ergebnis einer Panel-Diskussion unter zehn Expertenhöreren

effekten entgegenzuwirken. Ergebnis des Versuchs waren Merkmalsprofile für die 49 Interpretationen, welche zu jedem Interpretationsaspekt die Mittelwerte der zehn Expertenurteile wiedergeben und daher als quasi-objektive Eigenschaften der Beurteilungsgegenstände aufgefasst werden dürfen (Abbildung 1, umseitig).

Auf diese Weise wurden insgesamt 8170 Einzelbewertungen erhoben, 24 fehlende Werte wurden für die statistische Auswertung durch den Gruppenmittelwert ersetzt. Die stereofone Wiedergabe wurde mit qualitativ hochwertigen Aktivmonitoren (Klein & Hummel O300 D) im elektronischen Studio des Fachgebiets Audiokommunikation der TU Berlin realisiert, welches die Anforderungen an Abhörräume für Hörversuche nach ITU-R BS.1116-1 erfüllt.

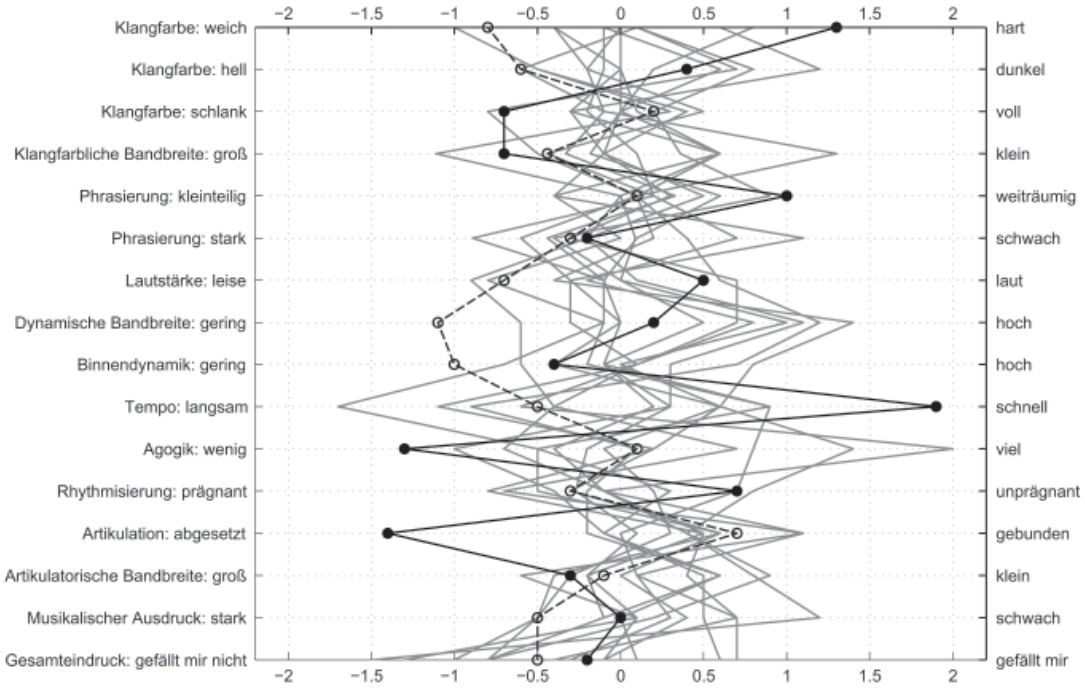

Abbildung 1: Merkmalsprofile für 16 Einspielungen der Klaviersonate F-Dur KV 322 von Wolfgang Amadeus Mozart, 1. Satz, T. 41–94. Hervorgehoben sind die Interpretationen 3 (Glenn Gould, aufgenommen 1972, ausgezogen) und 6 (Mitsuko Uchida, aufgenommen 1983, gestrichelt), die besonders unterschiedlich bewertet wurden.

3. Feature-Extraktion

Mit Hilfe einer an der TU Berlin entwickelten Bibliothek von Algorithmen zur Aufführungsanalyse[12] wurde zunächst eine Detektion der Notenanfänge in allen analysierten Einspielungen vorgenommen, falls erforderlich nach Gehör manuell korrigiert und durch einen Dynamic-Time-Warping-Algorithmus zu einem MIDI-kodierten Notentext synchronisiert. Anhand der identifizierten Notenanfänge wurden Inter-Onset-Intervalle (IOI) bzw. eine darauf beruhende Folge von Onset-bezogenen, lokalen Tempi in Beats per Minute (BPMons) berechnet. Um einen stärker geglätteten Verlauf des Makro-Tempos zu erhalten, wurden zusätzlich Inter-Bar-Intervalle (IBI) bzw. im Taktabstand errechnete BPM-Werte (BPMbar) berechnet, wofür jeweils die Zeitpunkte der ersten Zählzeit im Takt, falls im Notentext ausgelassen durch Interpolation ermittelt, herangezogen wurden.

[12] Vgl. auch Alexander Lerch, *Software Based Extraction of Objective Parameters from Music Performances,* München 2009.

Als Ergebnis des Synchronisationsvorgangs standen somit neben den unmittelbar daraus resultierenden Tempo-Werten drei Zeitraster für die Extraktion aller weiteren Features zur Verfügung:

- eine als ›Raw Grid‹ bezeichnete Folge von Abtastzeitpunkten, für die die sogenannte Hop Size maßgeblich ist, d. h. ein auf 10 ms festgelegter Zeitabstand, um den das für die Extraktion der Features erforderliche Zeitfenster bei der automatisierten Analyse verschoben wurde,
- ein als ›Event Grid‹ bezeichnetes Raster von Zeitpunkten für alle im Notentext ausgewiesenen musikalischen Einzelereignisse, und schließlich
- ein als ›Fixed Grid‹ bezeichnetes Raster von Zeitpunkten im Abstand von Sechzehntelnoten – unabhängig davon, ob diese durch musikalische Ereignisse besetzt sind oder nicht.

Während das ›Raw Grid‹ also der Realzeit der jeweiligen Interpretation folgt und somit auch bei jeder verschieden schnellen Einspielung eine unterschiedliche Länge hat, folgen das ›Event Grid‹ und das ›Fixed Grid‹ einer durch den Notentext vorgegebenen Schrittweite und weisen für alle Interpretationen die gleiche Länge auf.

Für jede der analysierten 49 Einspielungen und für jedes der drei beschriebenen Zeitraster wurde mit Hilfe der genannten Software eine Reihe von Features aus dem Audiosignal extrahiert, die zum Teil auf dem zeitlichen Verlauf der Signalamplitude und zum Teil auf spektralen Transformationen beruhen und physikalische Korrelate der perzeptiven Kategorien Klangfarbe und Lautheit bzw. Dynamik darstellen. Sie beinhalten eine Reihe von Intensitätsmaßen: von einem leistungsbezogenen, quadratischen Mittelwert der Signalamplitude (Root Mean Square, RMS) über typischerweise im Tonstudio verwendete Aussteuerungsmaße (Peak Programme Meter, VU Meter), einen in der akustischen Messtechnik häufig verwendeten frequenzbewerteten Schallpegel (dBA) und einen im Rundfunk eingesetzten Lautheitsprädiktor nach ITU-R BS.1770 bis zu komplexen Lautheitsmodellen nach Zwicker,[13] für die zwei Implementierungen nach DIN 45631:1991 und nach ITU-R BS.1387:2006 vorgenommen wurden.

Als klangfarblich relevante Merkmale wurden acht in der Audiosignalverarbeitung verbreitete spektrale Maße bestimmt. Sie messen den spektralen Schwerpunkt (Spectral Centroid), die spektrale Bandbreite um diesen Schwerpunkt (Spectral Spread), den Anteil hoher Frequenzen (Spectral Rolloff) und die Veränderlichkeit des Spektrums über die Zeit (Spectral Flux). Darüber hinaus wurden die ersten vier sogenannten Mel Frequency Cepstral Coefficients (MFCCs) berechnet, wie sie häufig bei der Klassifikation von Sprache und Musik eingesetzt werden. Da die Bedeutung der Features im Beitrag von Alexander Lerch in diesem Band genauer erläutert wird, soll hier nur eine tabellarische Übersicht gegeben werden (Tabelle 2, umseitig).

13 Hugo Fastl und Eberhard Zwicker, *Psychoacoustics. Facts and Models,* Berlin ³2010.

Kategorie	Feature	Label
Dynamik	Zwicker-Lautheit nach DIN 45631:1991	ZWDIN
	Zwicker-Lautheit nach ITU-R BS.1387:2006	ZW1387
	Quadratischer Mittelwert der Signalamplitude	RMS
	Lautheit nach ITU-R BS.1770	1770
	A-bewerteter Signalpegel	dBA
	Studiopegel (Peak Program Meter) nach DIN IEC 60268-10	PPM
	Studiopegel (VU Meter) nach DIN IEC 60268-17	VU
Klangfarbe	Spectral Flux	SF
	Spectral Rolloff	SR
	Spectral Centroid	SC
	Spectral Slope	SS
	Mel Frequency Cepstral Coefficient 0	MFCC0
	Mel Frequency Cepstral Coefficient 1	MFCC1
	Mel Frequency Cepstral Coefficient 2	MFCC2
	Mel Frequency Cepstral Coefficient 3	MFCC3
Tempo	Inter-Onset-Interval	IOI
	Inter-Bar-Interval	IBI
	Normalized Inter-Onset-Interval	IOInorm
	Beats Per Minute – bars	BPMbar
	Beats Per Minute – onsets	BPMons
	Beats Per Minute – bars (reciprocal avg.)	BPMbar_r
	Beats Per Minute – onsets (reciprocal avg.)	BPMons_r

Tabelle 2: In der verwendeten Softwarebibliothek zur Aufführungsanalyse implementierte Audio-Features. Die auf Dynamik und Klangfarbe bezogenen Features sind mit unterschiedlichen, für den jeweiligen Parameter typischen Zeitfenstern und einer Schrittweite (Hop Size) von 10 ms berechnet.

Während für die Lautheitsmaße unterschiedliche, jeweils geeignete Analysefenster benutzt wurden, sind alle spektralen Maße mit einer Fensterlänge von 23 ms, entsprechend 2^{10} Samples bei 44,1 kHz, berechnet.

Die resultierenden Zeitreihen der Features wurden im Folgenden getrennt für jedes der drei genannten Zeitraster ausgewertet. Für das ›Raw Grid‹ wurden alle extrahierten Zeitwerte, für das ›Event Grid‹ hingegen nur die im Analysefenster eines musikalischen Ereignisses befindlichen Werte verwendet. Für das ›Fixed Grid‹ wurden die auf einer linearen Zeitachse extrahierten Werte durch Interpolation mit kubischen Splines in einem partiturbezogenen Zeitraster im Abstand von Sechzehn-

Statistischer Deskriptor	**Label**
Arithmetisches Mittel	mu
Standardabweichung	std
Modus	md
Geometrisches Mittel	geom
0-Quantil	qu0
0,1-Quantil	qu10
0,25-Quantil	qu25
0,5-Quantil (Median)	qu50
0,75-Quantil	qu75
0,9-Quantil	qu90
1-Quantil	qu100
Variationsbreite $Q_1 - Q_0$	qu0100
Interdezilabstand $Q_{0,9} - Q_{0,1}$	qu1090
Interquartilabstand $Q_{0,75} - Q_{0,25}$	qu2575

Tabelle 3: Statistische Kennwerte zur Beschreibung der zentralen Tendenz und der Streuung aller extrahierten Feature-Zeitreihen

telnoten neu abgetastet. Im ersten Fall wurde also das gesamte Audiosignal analysiert, einschließlich der Pausen und der Zeiten zwischen den Notenanfängen. Im zweiten Fall wurden diese Lücken zwischen den Notenanfängen ignoriert, und im dritten Fall wurden die Notenzwischenräume zwar analysiert, aber durch ein an den Positionen der Noteneinsätze orientiertes Sechzehntel-Raster so diskretisiert, dass unabhängig vom musikalischen Tempo für jede Interpretation die gleiche Anzahl von Werten entsteht.

Um die analysierten Zeitreihen zu Einzahlwerten zusammenzufassen, die sich in Zusammenhang sowohl mit der mittleren Tendenz als auch mit der Veränderlichkeit von Interpretationseigenschaften bringen lassen (siehe Abschnitt 2), wurden verschiedene in der deskriptiven Statistik gebräuchliche Maße der zentralen Tendenz und der Streuung berechnet (Tabelle 3).

Dazu gehören Zentralwertmaße wie das arithmetische Mittel, das geometrische Mittel, der am häufigsten auftretende Wert (Modus) sowie sogenannte Quantilwerte, welche die Verteilung der Messwerte in anhand ihrer relativen Häufigkeit definierte Bereiche teilen, vom Minimum (0-Quantil) über den Median (0,5-Quantil) bis zum Maximum (1-Quantil). Als Streuungsmaße wurden neben der Standardabweichung verschiedene Interquantilabstände berechnet, die etwa die mittleren 50% der

Messwerte (Interquartilabstand, $Q_{0,75}$–$Q_{0,25}$), die mittleren 80% (Interdezilabstand, $Q_{0,9}$–$Q_{0,1}$) oder den gesamten Wertebereich (Variationsbreite) markieren.

Da im Hörversuch innerhalb desselben Werks nur relative Urteile zu den Interpretationen abgegeben wurden, etwa über die gespielten Tempi (siehe Abschnitt 2), mussten auch die technisch ermittelten Maße so transformiert werden, dass sie nur noch die Abweichung einer Interpretation vom Mittelwert aller Interpretationen desselben Werks angeben. Diese Bereinigung um die werkspezifischen absoluten Werte wurde durch eine für jedes Werk getrennte z-Transformation erreicht. Dadurch erhält zum Beispiel eine Interpretation mit durchschnittlichem Tempo den Wert 0, relativ langsamere oder schnellere Einspielungen erhalten negative oder positive Werte, so dass die Standardabweichung aller Werte 1 beträgt.

Die z-transformierten (standardisierten) Merkmale wurden in einem letzten Schritt drei verschiedenen Transformationen unterzogen, um verschiedene funktionale Zusammenhänge zwischen den signalbasierten Variablen VAR und den Beurteilungen der Hörer zu untersuchen. Bei der linearen Transformation Lin_VAR wurden die Werte unverändert übernommen für den Fall, dass für jede analysierte Interpretation n gleichabständig größere Feature-Werte gleichabständig größeren Beurteilungswerten entsprechen.

$$\text{Lin\_VAR}(n) = \text{VAR}(n)$$

Durch eine logarithmische Transformation Log_VAR wird erreicht, dass nicht Abstände sondern Größenverhältnisse bei den technisch ermittelten Werten in entsprechende potentielle Empfindungswerte übersetzt werden. Dies ist durch das Webersche Gesetz der Psychologie motiviert, nach dem ebenmerkliche Unterschiede einer Empfindungsgröße durch ein bestimmtes Verhältnis der Reizstärken, nicht durch deren physikalisch gemessene Differenz gegeben sind.

$$\text{Log\_VAR}(n) = \log 10\,(a + 1 + \text{Lin\_VAR}(n))$$

Da der Logarithmus nur für positive Zahlen definiert ist und nur der positive Abschnitt der Funktion verwendet werden sollte, mussten die z-transformierten Werte vor der funktionalen Transformation um (a+1) nach rechts verschoben werden, wobei für a das Minimum der z-transformierten Werte aller Features und aller Interpretationen eingesetzt wurde.

Schließlich wurde eine negativ quadratische Transformation x^2_VAR durchgeführt. Sie ist durch die Beobachtung motiviert, dass bei evaluativen ästhetischen Urteilen häufig ein Optimum vorliegt, wenn bestimmte Eigenschaften eines Reizes einen bestimmten Wert weder unter- noch überschreiten.

$$\text{x\^{}2\_VAR}\,(n) = -(\text{VAR}(n))^2$$

Da die im folgenden vorgenommenen statistischen Zusammenhangsanalysen (bivariate Korrelation, multiple Regression) nur lineare Zusammenhänge identifizieren können, erlaubt es die vorweggenommene Transformation der unabhängigen Variable (akustische Features), auch andere, in diesem Fall also logarithmische und um-

gekehrt quadratische, Zusammenhänge zwischen den akustischen und den wahrgenommenen Eigenschaften von Interpretationen zu ermitteln. Abbildung 2 illustriert die drei untersuchten funktionalen Zusammenhänge.

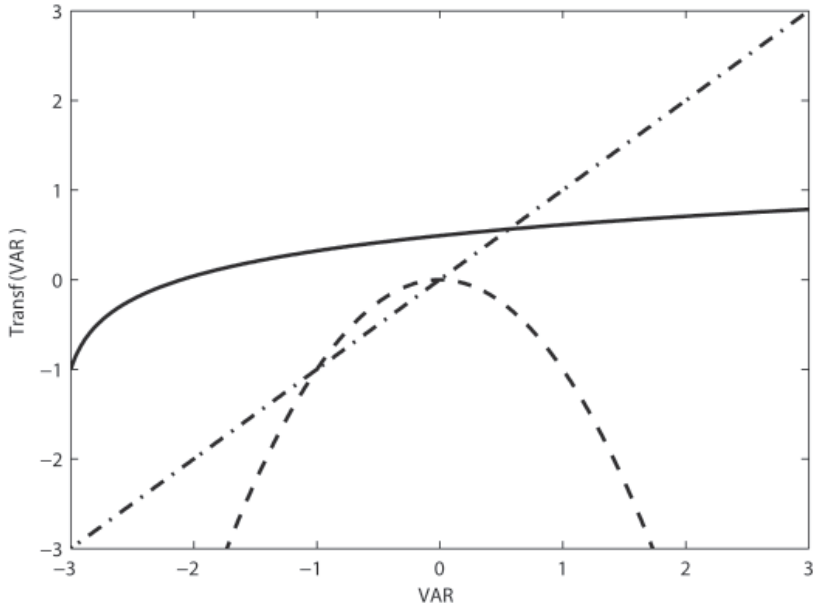

Abbildung 2: Drei funktionale Transformationen, die prototypischen Zusammenhängen zwischen den Verteilungen von signalbasierten und perzeptiven Merkmalen Rechnung tragen: linear (strichpunktiert), logarithmisch (ausgezogen) und umgekehrt quadratisch (gestrichelt).

Insgesamt wurden somit für 18 technische Merkmale, 3 verschiedene Zeitraster, 14 verschiedene statistische Kennwerte und 3 psychophysische Transformationen insgesamt mehr als 2000 Deskriptoren berechnet. Zwischen diesen für jede Interpretation technisch extrahierten Deskriptoren und den Beurteilungen der Expertenhörer wurden im Folgenden zwei Arten statistischer Zusammenhänge analysiert. Zunächst wurde derjenige technische Einzeldeskriptor identifiziert, der am höchsten mit einem einzelnen Merkmal des semantischen Differentials (Tabelle 1) korreliert. In einem zweiten Schritt wurde anhand von Regressionsmodellen überprüft, inwieweit Urteile über die Spezifika der untersuchten Interpretationen durch eine Kombination mehrerer technischer Deskriptoren erklärt werden können.

4. Ergebnisse

Inwieweit die Expertenhörer die im semantischen Differential enthaltenen Attribute (Tabelle 1) übereinstimmend (konkordant und konsistent) verwendet hatten, wurde anhand der Intra-Klassen-Korrelation (ICC Typ 3,[14] berechnet für alle Attribute, Urteiler, Musikstücke und Interpretationen) ermittelt, die auch als Intersubjektivitäts- bzw. Objektivitätsmaß gelten kann. Die Urteilsübereinstimmung des gesamten Panels (average) ist für fast alle Attribute ausreichend ($ICC_{adj}(2,k) \geq 0{,}7$) oder gut ($ICC_{adj}(2,k) \geq 0{,}8$). Lediglich die Attribute zur Bewertung der Phrasierung (schwach–stark und kleinteilig–weiträumig) wurden von den Hörern offensichtlich nicht übereinstimmend verwendet. Die Urteilsübereinstimmung über das gesamte Mess-

	$ICC_{adj}(2,k)$
Klangfarbe (weich–hart)	0,829
Klangfarbe (dunkel–hell)	0,802
Klangfarbe (schlank–voll)	0,687
Klangfarbliche Bandbreite (klein–groß)	0,777
Phrasierung (kleinteilig–weiträumig)	0,306
Phrasierung (schwach–stark)	0,587
Lautstärke (leise–laut)	0,776
Dynamik (gering–hoch)	0,870
Binnendynamik (gering–hoch)	0,699
Tempo (langsam–schnell)	0,962
Agogik (wenig–viel)	0,898
Rhythmisierung (unprägnant–prägnant)	0,702
Artikulation (gebunden–abgesetzt)	0,832
Artikulatorische Bandbreite (klein–groß)	0,765
Musikalischer Ausdruck (schwach–stark)	0,773
Gesamteindruck (gefällt nicht–gefällt)	0,761
Messinstrument (alle Attribute)	0,793

Tabelle 4: Adjustierte Intra-Klassen-Koeffizienten (ICCs) als Maß für die Urteilsübereinstimmung bei der Anwendung des semantischen Differentials (Tabelle 1) durch das gesamte Urteiler-Panel (average)

[14] Patrick E. Shrout und Joseph L. Fleiss, ›Intraclass Correlations: Uses in Assessing Rater Reliability‹, in: *Psychological Bulletin* 86/2 (1979), S. 420–428.

instrument, bestehend aus Fragebogen und Urteiler-Panel, ist ausreichend bis gut ($ICC_{adj}(2,k) = 0{,}793$; Tabelle 4). Die durch eine multivariate Varianzanalyse (MANOVA) nachgewiesenen großen und hochsignifikanten multivariaten Mittelwertunterschiede zwischen den untersuchten Interpretationen ($p < 0{,}0005$; partial $\eta^2 = 0{,}321$; $1-\beta = 1{,}0$), sind darüber hinaus ein Indikator für die Gültigkeit des Messinstruments zur Charakterisierung der Interpretationen.

Um den Zusammenhang zwischen dem Höreindruck und den Signaleigenschaften zu untersuchen, wurde zunächst die Produkt-Moment-Korrelation zwischen jedem signalbasierten Deskriptor und jedem perzeptiven Merkmal berechnet.

Die Korrelationen dieser Einzeldeskriptoren (Tabelle 5) nehmen Werte zwischen 0,4 und 0,96 an. Nach dem Kriterium, dass mehr als die Hälfte der Varianz der ab-

Abhängige Variable	**Bester Einzeldeskriptor**	***r***	***r²***
Klangfarbe (weich–hart)	Lin_SR_mu	0,751	0,564
Klangfarbe (dunkel–hell)	Lin_SR_qu75	0,673	0,453
Klangfarbe (schlank–voll)	Log_IBI_qu90	0,561	0,315
Klangfarbliche Bandbreite (klein–groß)	Log_PPM_qu10	–0,620	0,384
Phrasierung (kleinteilig–weiträumig)	Lin_IOI_qu2575	–0,425	0,181
Phrasierung (schwach–stark)	Lin_IOInorm_std	0,480	0,230
Lautstärke (leise–laut)	Log_SR_geom	0,529	0,280
Dynamik (gering–hoch)	Lin_ZWDIN_qu1090	0,655	0,429
Binnendynamik (gering–hoch)	Lin_MFCC2_md	–0,506	0,256
Tempo (langsam–schnell)	Log_BPMbar_mu	0,963	0,927
Agogik (wenig–viel)	Log_BPMbarR_std	–0,696	0,484
Rhythmisierung (unprägnant–prägnant)	Log_BPMonsR_md	0,562	0,316
Artikulation (gebunden–abgesetzt)	Log_IOI_qu25	0,498	0,248
Artikulatorische Bandbreite (klein–groß)	x^2_SS_qu25	0,404	0,163
Musikalischer Ausdruck (schwach–stark)	Log_BPMonsR_qu1090	–0,524	0,275
Gesamteindruck (gefällt nicht–gefällt)	Lin_SS_qu10	–0,498	0,248

Tabelle 5: Der beste signalbasierte Einzeldeskriptor (höchste Produkt-Moment-Korrelation r bzw. Varianzaufklärung r^2) für jedes der 16 perzeptiv bewerteten Merkmale von musikalischen Interpretationen. Das Label für den Deskriptor setzt sich aus den Bezeichnungen für die vor der Korrelation ausgeführte funktionale Transformation, dem Feature-Label (Tabelle 2) und dem statistischen Kennwert (Tabelle 3) zusammen.

hängigen Variable (Hörurteile) durch den technischen Deskriptor erklärt werden kann ($r^2 > 0{,}5$), gibt es nur zwei Attribute, die durch einen Einzeldeskriptor gut erklärt werden können. Das ist zum einen das mittlere Tempo, das fast vollständig ($r^2 = 0{,}93$) dem arithmetischen Mittel der taktweise erhobenen BPM-Werte entlang eines logarithmischen Verlaufs (Log_BPMbar_mu) folgt. Dies bestätigt die bereits von Repp in einem Hörversuch gemachte Beobachtung, dass das arithmetische Mittel der Tempi das empfundene mittlere Tempo besser erklärt als jeder Quantilwert der Tempoverteilung,[15] solange keine starken Ritardandi oder Tempowechsel auftreten, was auch in unseren Ausschnitten nicht der Fall war. Zum anderen kann die Klangfarbe entlang einer Skala von weich bis hart gut ($r^2 = 0{,}56$) durch den Mittelwert des Spectral Rolloff erklärt werden. Für alle weiteren auf Klangfarbe, Dynamik und Tempo bezogenen Attribute können einzelne Deskriptoren immerhin zwischen 25 und 50% der Varianz erklären. Lediglich für komplexere musikalische Konzepte wie Artikulation, Phrasierung und den ästhetischen Gesamteindruck können einzelne akustische Parameter keinen nennenswerten Beitrag zur Erklärung der Hörerurteile leisten.

Nicht alle Zusammenhänge sind leicht zu interpretieren. Zwar ist es unmittelbar plausibel, dass die dynamische Bandbreite der Interpretation durch die Variationsbreite eines Lautheitsmaßes (Interdezilabstand der Zwicker-Lautheit nach DIN) und die Agogik durch die Variationsbreite eines Tempodeskriptors (Standardabweichung der reziprok gemittelten BPM-Werte) erklärt werden kann. Andererseits mag es auf den ersten Blick überraschend sein, dass sich als bester Einzeldeskriptor für die Klangfarbe in der Dimension ›schlank–voll‹ das 0,1-Quantil des Inter-Bar-Intervalls erweisen hat. Auf den zweiten Blick ist es allerdings keineswegs unplausibel, dass die Entfaltung eines ›vollen‹ Tons eine gewisse Tondauer erfordert, dass also ein höheres Tempo mit der Wahrnehmung eines ›schlankeren‹ Tons korreliert.

Dass die Korrelation zwischen den akustischen und den wahrgenommenen Eigenschaften der untersuchten Interpretationen in vielen Fällen unbefriedigend ist, muss allerdings nicht zwangsläufig auf die begrenzte Validität der berechneten Deskriptoren zurückzuführen sein. Es ist vielmehr zu erwarten, dass sich zahlreiche der wahrgenommenen Qualitäten nur durch eine Kombination verschiedener technischer Merkmale gut erklären lassen, dass also zum Beispiel die Klangfarbe immer dann als ›schlank‹ empfunden wird, wenn ein hohes Tempo mit einer niedrigen Lautstärke zusammenfällt. Um das Erklärungspotential solcher kombinierter Modelle zu überprüfen, wurde eine multiple Regressionsanalyse durchgeführt. Hierbei werden lineare Gleichungssysteme so konstruiert, dass sich ein perzeptives Merkmal möglichst gut durch eine Kombination verschiedener technischer Deskriptoren ergibt.

Auf Basis der vorliegenden Daten zu 49 Einzelfällen (Interpretationen) lassen sich komplexe Modelle mit bis zu 48 unabhängigen Variablen bilden. Hierbei wurden für jede Regressionsanalyse die 48 Deskriptoren mit der höchsten Varianzauf-

15 Repp, ›On Determining the Basic Tempo‹ (wie Anm. 6).

klärung des perzeptiven Merkmals vorausgewählt. In den meisten Fällen leisten allerdings nur einige dieser Deskriptoren einen nennenswerten Beitrag zur Erklärung des perzeptiven Merkmals als abhängige Variable, zumal wenn sie untereinander bereits hoch korreliert sind. Im vorliegenden Fall wurden Deskriptoren nur dann in das Regressionsmodell aufgenommen, wenn sie zu einer statistisch signifikanten Zunahme der erklärten Varianz führen (stepwise procedure).[16] Dieses Selektionsverfahren kann nicht gewährleisten, dass das optimale Modell gefunden wird, welches mit einer minimalen Anzahl von Deskriptoren eine maximale Varianzaufklärung leistet, da sich sowohl die getroffene Vorauswahl der Deskriptoren als auch die Reihenfolge, mit der diese in die Regression einbezogen werden, auf die Struktur des resultierenden Modells auswirken kann, und da die Modellkomplexität nicht als Aufnahme- bzw. Ausschlusskriterium berücksichtigt wird. Allerdings kann im vorliegenden Fall nicht auf spezifisches Vorwissen zurückgegriffen werden, das eine bessere Vorauswahl von Deskriptoren erlaubt hätte als die vorliegende, nach einem rein numerischen Kriterium getroffene. Im Hinblick auf eine explorative Untersuchung der Zusammenhänge und der Größenordnung, mit der komplexe Merkmale musikalischer Interpretationen in einem an den empirischen Daten optimierten Modell durch eine elementare Signalanalyse erklärt werden können, erscheint das gewählte Verfahren jedoch angemessen.

Generell zeigen diejenigen Modelle, die Deskriptoren enthalten, die auf im ›Event Grid‹ gerasterten Features basieren, die höchste Varianzaufklärung der Daten. Der Mittelwert der schrumpfungskorrigierten Determinationskoeffizienten beträgt $R^2_{adj} = 54\%$, während er für das ›Raw Grid‹ bei $R^2_{adj} = 51\%$ und für das ›Fixed Grid‹ bei $R^2_{adj} = 49\%$ liegt. Offenbar mindern Features, die nicht auf Noteneinsätze, sondern auf Tondauern und Pausen bezogen sind, das perzeptive Erklärungspotential der Deskriptoren. Daher werden im Folgenden nur Ergebnisse betrachtet, die auf den im ›Event Grid‹ gerasterten Features beruhen.

Folgende Voraussetzungen der multiplen Regressionsanalyse wurden überprüft: 1. Aufgrund der automatischen Feature-Extraktion gibt es keine fehlenden Werte. 2. Mit einer Ausnahme liegen die Cook's Distances unterhalb der kritischen Werte von 0,85 für vier, 0,8 für drei und 0,7 für zwei Deskriptoren[17] und zeigen an, dass in den Daten keine relevanten Ausreißer vorliegen. 3. Kolmogorov-Smirnov-Lilliefors-Tests zeigen, dass alle Residuen ausreichend normalverteilt sind. 4. Die Autokorrelation der Residuen wurde mit Hilfe von Durbin-Watson-Tests und durch visuelle Inspektion der zugehörigen Scatterplots überprüft. 5. Bei Betrachtung der Scatterplots der studentisierten Residuen gegen die standardisierten abhängigen Variablen erwiesen sich die Varianzen als ausreichend homogen (Homoskedastizität). 6. Die

[16] Hierfür wurde mit einem partiellen F-Test gegen die Nullhypothese getestet, dass die Zunahme der Varianzaufklärung gleich Null ist. Das Signifikanzniveau für die Einbeziehung und die Entfernung einer Variable wurde auf $p = 0{,}05$ und $p = 0{,}10$ festgesetzt, vgl. Jürgen Janssen und Wilfried Laatz, *Statistische Datenanalyse mit SPSS*, Heidelberg u. a. [7]2010.

[17] Barry McDonald, ›A Teaching Note on Cook's Distance – A Guideline‹, in: *Research Letters in the Information and Mathematical Sciences* 3 (2002), S. 127–128.

gegenseitige lineare Unabhängigkeit der Deskriptoren ist bis auf vier Ausnahmen ausreichend.[18]

Im Ergebnis liefert die regressionsanalytische Auswertung Modelle, welche die von den Hörern bestimmten Eigenschaften der 49 Interpretationen durch eine Kombination von ein bis sechs signalbasierten Deskriptoren statistisch erklären können. Danach kann für zehn der sechzehn Merkmale mehr als die Hälfte der von den Hörern wahrgenommen Unterschiede zwischen den Interpretationen desselben Stücks durch die extrahierten akustischen Merkmale erklärt werden, wenn man diese kombiniert. Dazu gehören alle auf Tempo, Klangfarbe und Intensität bezogenen Interpretationsmerkmale, ebenso wie die evaluativen Urteile zur Stärke des musikalischen Ausdrucks und zum Gesamteindruck (Abbildung 3).

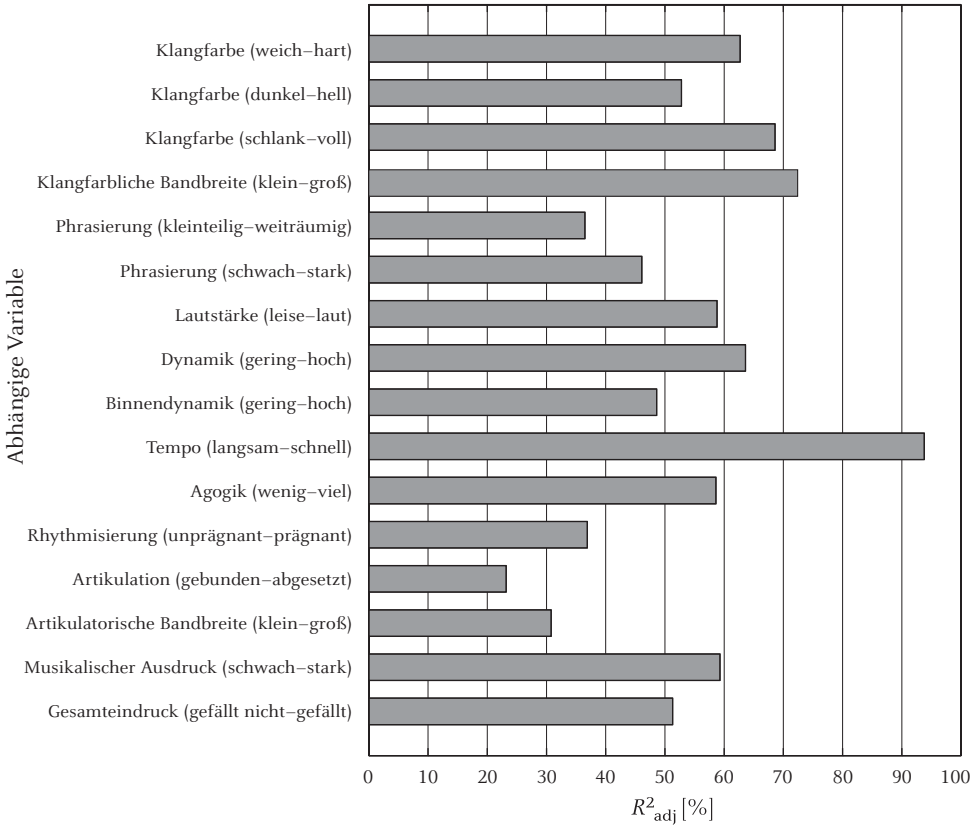

Abbildung 3: Durch signalbasierte Deskriptoren erklärte schrumpfungskorrigierte Varianz verschiedener wahrgenommener Interpretationsmerkmale

[18] Bestimmt wurde die Toleranz $1–R^2$, die für 50 der insgesamt 54 Deskriptoren ausreichend hohe, für je zwei Deskriptoren in den Modellen Klangfarbliche Bandbreite und Rhythmisierung hingegen kritische Werte unter 0,1 annimmt (vgl. Markus Bühner und Matthias Ziegler, *Statistik für Psychologen und Sozialwissenschaftler*, München 2009, S. 678).

Abhängige Variable	R^2	Adj. R^2	F	Deskriptoren	Stand. Koeff. Beta	t	Sig. p
Klangfarbe (weich–hart)	0,651	0,627	27,941	Constant			
				Lin_ZWDIN_qu50	0,214	2,169	0,035
				Lin_MFCC1_qu75	0,568	2,779	0,008
				Lin_SR_mu	1,168	5,717	0,000
Klangfarbe (dunkel–hell)	0,548	0,528	27,830	Constant			
				Lin_MFCC3_qu25	–0,311	–3,11	0,003
				Lin_SR_qu75	0,631	6,304	0,000
Klangfarbe (schlank–voll)	0,725	0,686	18,459	Constant			
				Log_IOI_qu75	–0,718	–2,89	0,006
				Log_BPMons_md	–0,354	–2,65	0,011
				Log_MFCC3_md	0,35	4,02	0,000
				Lin_SR_qu90	–0,274	–2,89	0,006
				x^2_MFCC1_qu75	0,346	3,961	0,000
				Lin_BPMbar_qu10	–0,651	–2,68	0,011
Klangfarbliche Bandbreite (klein–groß)	0,752	0,724	26,12	Constant			
				Lin_dBA_qu_10	0,616	2,868	0,006
				Log_ZW1387_qu1090	1,869	3,963	0,000
				Lin_ZW1387_qu1090	–1,676	–3,42	0,001
				Log_IOI_norm_std	0,397	4,895	0,000
				Log_PPM_qu_10	–1,192	–5,37	0,000
Phrasierung (kleinteilig–weiträumig)	0,405	0,365	10,194	Constant			
				Log_SC_qu90	–0,377	–3,24	0,002
				x^2_VU_qu2575	–0,349	–2,95	0,005
				Lin_IOI_qu2575	–0,34	–2,89	0,006
Phrasierung (schwach–stark)	0,506	0,461	11,253	Constant			
				Lin_SF_qu50	–0,26	–2,43	0,019
				x^2_BPMons_qu90	0,346	3,238	0,002
				Log_VU_qu2575	0,3	2,8	0,008
				Log_BPMons_qu1090	0,446	4,178	0,000
Lautstärke (leise–laut)	0,614	0,588	23,862	Constant			
				Log_ZWDIN_mu	0,459	4,716	0,000
				Lin_SF_qu10	–0,508	–5,28	0,000
				x^2_dBA_md	–0,444	–4,73	0,000

Tabelle 6: Multiple Regressionsmodelle für die Erklärung der 16 von Expertenhörern beurteilten Merkmale musikalischer Interpretation durch signalbasierte Deskriptoren. Durch die Stepwise-Prozedur wurden nur Deskriptoren in das Modell aufgenommen, die einen signifikanten Beitrag zur Erklärung der abhängigen Variable liefern (rechte Spalte). Die Signifikanz des Gesamtmodells gemäß F-Test ist in allen Fällen sehr hoch ($p < 0{,}0005$).

Abhängige Variable	R^2	Adj. R^2	F	Deskriptoren	Stand. Koeff. Beta	t	Sig. p
Dynamik (gering–hoch)	0,659	0,636	28,973	Constant			
				x^2_SS_qu25	0,31	3,368	0,002
				Log_ZW1387_qu0100	0,476	5,426	0,000
				Lin_PPM_qu10	–0,447	–4,82	0,000
Binnendynamik (gering–hoch)	0,529	0,486	12,338	Constant			
				Lin_MFCC0_md	–0,306	–2,85	0,007
				x^2_VU_qu10	0,283	2,564	0,014
				x^2_MFCC1_qu25	0,315	2,871	0,006
				Lin_MFCC2_md	–0,283	–2,51	0,016
Tempo (langsam–schnell)	0,940	0,938	361,756	Constant			
				Log_BPMons_qu100	0,166	3,043	0,004
				Log_BPMbar_mu	0,839	15,36	0,000
Agogik (wenig–viel)	0,620	0,586	17,975	Constant			
				Log_SF_qu0	0,348	3,633	0,001
				x^2_IOI_qu10	0,319	3,164	0,003
				Lin_BPMbar_qu0	–0,239	–2,17	0,035
				Lin_BPMbar_std	0,382	3,632	0,001
Rhythmisierung (unprägnant–prägnant)	0,395	0,369	15,044	Constant			
				Log_BPMons_qu75	–1,296	–2,48	0,017
				Log_BPMons_qu50	1,826	3,487	0,001
Artikulation (gebunden–abgesetzt)	0,248	0,232	15,474	Constant			
				Log_IOI_qu25	0,498	3,934	0,000
Artikulatorische Bandbreite (klein–groß)	0,351	0,308	8,124	Constant			
				Log_BPMbar_std	0,317	2,629	0,012
				Lin_MFCC0_md	–0,334	–2,72	0,009
				x^2_SS_qu25	0,337	2,755	0,008
Musikalischer Ausdruck (schwach–stark)	0,635	0,593	14,967	Constant			
				Lin_BPMbar_qu1090	0,39	4,214	0,000
				x^2_MFCC1_qu25	0,252	2,615	0,012
				Log_ZW1387_qu1090	0,716	3,262	0,002
				Lin_ZWDIN_qu1090	–0,744	–2,32	0,025
				Log_ZW1387_qu10	–0,86	–3,86	0,000
Gesamteindruck (gefällt nicht–gefällt)	0,554	0,513	13,654	Constant			
				x^2_VU_qu2575	–0,235	–2,22	0,032
				Lin_SC_qu100	–0,354	–3,42	0,001
				Log_ZW1387_qu0	–0,352	–3,36	0,002
				x^2_ZWDIN_mu	0,422	4,081	0,000

Tabelle 6 (Fortsetzung)

Eine genaue Betrachtung der Zusammenhänge zwischen wahrgenommenen Qualitäten und akustischen Parametern für die Kategorien Tempo, Dynamik, Klangfarbe, Artikulation und Phrasierung liefert ein differenzierteres Bild.

Das von den Hörern beurteilte Tempo, das bereits durch die taktweise erhobenen und arithmetisch gemittelten BPM-Werte (Log_BPMbar_mu) alleine sehr gut beschrieben werden kann, lässt sich noch geringfügig besser erklären, wenn man zusätzlich das Maximum des über musikalische Einzelereignisse erhobenen Tempos mit einbezieht. Auch in einem kombinierten Modell liefern die BPM-Werte, die bereits durch die hohe bivariate Determination von 93% als hervorragender Deskriptor ausgewiesen sind, den größten Beitrag, was an den standardisierten β-Werten im Regressionsmodell abgelesen werden kann. In der untersuchten Stichprobe tragen aber offensichtlich auch einzelne, stark beschleunigte Passagen, die einen hohen Maximalwert der BPM-Verteilung (Log_BPMons_qu100) erzeugen, geringfügig zur Wahrnehmung eines hohen Tempos bei. Die Wahrnehmung unterschiedlicher Tempi folgt in beiden Fällen einer logarithmischen BPM-Kurve, d. h. der Unterschied zwischen M. M. = 50 und M. M. = 60 wird offensichtlich so groß bewertet wie zwischen M. M. = 100 und M. M. = 120. Zur Wahrnehmung einer hohen Agogik trägt nicht nur eine hohe Streuung der taktweise erhobenen BPM-Werte bei (Lin_BPMbar_std), sondern auch das Ausmaß, in dem einzelne Takte gegenüber dem mitteleren Tempo verlangsamt werden, d. h. das untere Ende der BPM-Verteilung (Lin_BPMbar_qu0).

Die Modelle für Klangfarbe enthalten jeweils eine Mischung von auf spektralen und energetischen Features basierenden Deskriptoren und erklären zwischen 53% und 72% der wahrgenommenen Unterschiede. Während die Aspekte ›Helligkeit‹ (dunkel–hell) und ›Härte‹ (weich–hart) überwiegend durch den Mittelwert bzw. das 0,75-Quantil des Spectral Rolloff erklärt werden, ergibt sich die ›Tonfülle‹ (schlank–voll) aus einer Kombination von auf Spektrum und Zeitintervalle bezogenen Deskriptoren. Ein voller Ton ist somit nicht nur mit einer Übergewichtung tieffrequenter spektraler Anteile, sondern auch mit einem nicht zu raschen Tempo korreliert, wenngleich aber auch mit einem nicht zu großen Zeitabstand zwischen den einzelnen Noteneinsätzen. Das Modell für klangfarbliche Bandbreite wird überwiegend durch Streumaße der prädizierten Lautheit bestimmt, enthält aber ebenfalls einen Tempo-Deskriptor. Bemerkenswert ist die hohe Varianzaufklärung von $R^2_{adj} = 72\,\%$.

Die Modelle für Lautstärke und dynamische Bandbreite enthalten eine Kombination von Deskriptoren, die auf Features für prädizierte Lautheit und Spektrum basieren, und erklären zwischen 59% und 64% der Varianz. Während für die absolute Lautstärke – worauf schon zu Beginn hingewiesen wurde – schwer zu unterscheiden ist, ob die im Modell enthaltenen Features (SF, ZWDIN, dBA) Defizite in der Kompensation der Aussteuerung nach ITU-R BS.1770 erklären oder eine über die Signaleigenschaften hinausgehende, mittlere ›musikalische‹ Dynamikstufe, kann die dynamische Bandbreite durch eine hohe Variationsbreite der Lautheit (Log_ZW1387_qu0100), eine starke Zurücknahme der Lautstärke in leisen Passagen (Lin_PPM_qu10) und eine optimale spektrale Breite (Spectral Spread) gut erklärt werden ($R^2_{adj} = 64\,\%$). Der Eindruck einer hohen Binnendynamik scheint überwie-

gend durch spektrale Eigenschaften erklärbar zu sein (MFCC0, MFCC1, MFCC2), ohne dass der Zusammhang im Detail leicht zu interpretieren wäre.

Die auf Artikulation, Rhythmisierung und Phrasierung bezogenen Attribute werden durch die berechneten Deskriptoren nur zu einem geringen Anteil erklärt. Immerhin korrespondiert die Stärke der Phrasierung gut ($R^2_{adj} = 46\,\%$) mit einer hohen Variation von Tempo (Log_BPMons_qu1090) und Signalpegel (Log_VU_qu2575) als Mittel zur Betonung von Phrasierungsgrenzen, während die Bedeutung von zwei weiteren Parametern zunächst schwer zu interpretieren ist. Da keine Features zur Tondauer (für Artikulation bzw. artikulatorische Bandbreite) oder zur zeitlichen Struktur von Tempo- und Lautheitsmodulationen (für Phrasierungsweite und Rhythmisierung) erhoben wurden, ist die geringe Aufklärung dieser Interpretationsmerkmale nicht verwunderlich.

In Anbetracht dieser Defizite bei der Erfassung wichtiger struktureller Eigenschaften der untersuchten Interpretationen ist es überraschend, wie gut dennoch der Gesamteindruck ($R^2_{adj} = 51\,\%$) und die empfundene Stärke des musikalischen Ausdrucks ($R^2_{adj} = 59\,\%$) durch eine elementare Signalanalyse erklärt werden können. Hierbei korreliert Ausdrucksstärke vor allem mit einer hohen Variationsbreite von Lautstärke (Log_ZW1387_qu1090) und Tempo (Lin_BPMbar_qu1090), während ein positiver Gesamteindruck in der vorliegenden Stichprobe überwiegend auf eine optimale prädizierte Lautheit sowie auf eine geringe Maximalfrequenz des (veränderlichen) energetischen Schwerpunkts zurückgeht, wenngleich das Zusammenwirken der Deskriptoren aufgrund der im vorliegenden Versuch nicht auflösbaren Interaktion von ›musikalischer Lautstärke‹, technischer Aussteuerung und deren Kompensation schwer interpretierbar ist.

Instruktiv im Hinblick auf die Leistung der signalbasierten Deskriptoren erscheint jedoch eine Analyse, inwieweit die beiden evaluativen Gesamteinschätzungen ›musikalischer Ausdruck‹ und ›Gesamteindruck‹ durch eine Kombination der übrigen Merkmale, die im Zusammenhang mit tieferliegenden perzeptiven Verarbeitungsstufen stehen, erklärt werden können (Tabelle 7). Hierfür wurde ebenfalls eine Regressionsanalyse gemäß der Stepwise-Prozedur durchgeführt, mit denselben Kriterien für die Aufnahme und die Entfernung eines Deskriptors aus dem Modell. Alle Voraussetzungen für die Modellierung (s. o.) waren erfüllt.

Die Ergebnisse der Regressionsanalysen zeigen, dass die Bewertungen von musikalischem Ausdruck und ›Gesamteindruck‹ durch die perzeptiven Merkmale nur geringfügig besser erklärt werden können ($R^2_{adj} = 67\,\%$ und $R^2_{adj} = 55\,\%$) als durch die signalbezogenen Deskriptoren. Die Wahrnehmung eines starken musikalischen Ausdrucks ist hier verbunden mit hoher Agogik, starker Phrasierung, hoher klangfarblicher Bandbreite und weiträumiger Phrasierung (in der Reihenfolge ihres Beitrags zur Erklärung der abhängigen Variable), während der Gesamteindruck im Sinne einer ästhetischen Präferenz der Urteilergruppe mit weiträumiger Phrasierung, vollem Ton, starker Phrasierung und prägnanter Rhythmisierung (ebenfalls in der Reihenfolge ihres Beitrags) verbunden ist.

Abhängige Variable	R^2	Adj. R^2	F	Deskriptoren	Stand. Koeff. Beta	t	Sig. p
Musikalischer Ausdruck (schwach–stark)	0,699	0,671	25,529	Constant		2,760	0,008
				Agogik (wenig–viel)	0,378	3,397	0,001
				Phrasierung (schwach–stark)	0,344	3,305	0,002
				Klangfarbliche Bandbreite (klein–groß)	0,331	3,189	0,003
				Phrasierung (kleinteilig–weiträumig)	0,258	2,721	0,009
Gesamteindruck (gefällt nicht–gefällt)	0,589	0,552	15,781	Constant		–6,437	0,000
				Phrasierung (kleinteilig–weiträumig)	0,592	5,689	0,000
				Klangfarbe (schlank–voll)	0,395	3,932	0,000
				Phrasierung (schwach–stark)	0,328	3,042	0,004
				Rhythmisierung (unprägnant–prägnant)	0,263	2,510	0,016

Tabelle 7: Multiple Regressionsanalyse: Perzeptive Deskriptoren für die Bewertung des musikalischen Ausdrucks und des Gesamteindrucks

5. Diskussion

In der vorliegenden Studie wurde untersucht, inwieweit globale Ausdrucksqualitäten von musikalischen Interpretationen durch signalbezogene akustische Features erklärt werden können, die ihrerseits Modelle für elementare sensorische Eindrücke wie Lautheit oder Klangfarbe enthalten. Insbesondere wurde dabei ermittelt, welches Zeitraster (auf die Zeit, das musikalische Ereignis oder den musikalischen Schlag bezogen), welche funktionalen Transformationen und welche statistischen Kennwerte des Zentralwerts und der Streuung bei der Berechnung der akustischen Deskriptoren für die Erklärung der perzeptiven Interpretationsmerkmale geeignet sind.

Hier zeigt sich zunächst, dass die unterschiedlichen wahrgenommenen Tempi der 49 Interpretationen fast perfekt durch taktweise erhobene und arithmetisch gemittelte BPM-Werte erklärt werden können ($R^2_{adj} = 94\%$), während die empfundene Agogik durch die Streuung der BPM-Werte und verschiedene Verteilungsmaße für Tempo und Spektrum immer noch gut erklärt werden kann ($R^2_{adj} = 59\%$). Hier wäre durch Deskriptoren, die nicht nur auf die Verteilung, sondern auch auf den zeitlichen Verlauf der Tempogestaltung Bezug nehmen, vermutlich eine bessere Erklärung möglich, wenn man bedenkt, dass der Unterschied zwischen einem ständig modulierten Binnentempo und einem terrassenförmig angelegten Tempoverlauf, der Tempowechsel nur an strukturellen Zäsuren vornimmt, durch einen rein statistischen Deskriptor nicht erfasst werden, obwohl beide Tempoverlaufsformen sicher in ihrer agogischen Wirkung unterschiedlich wahrgenommen werden.

Im Hinblick auf eine Erklärung der Eindrücke von Lautheit und Klangfarbe erweist sich ein auf musikalische Ereignisse bezogenes Zeitraster, das Features nur im zeitlichen Kontext von Notenanfängen extrahiert, als besonders leistungsfähig, mehr als eine Analyse des gesamten Audiosignals, sei es in Echtzeit oder durch ein partiturbezogenes Raster abgetastet. Mit diesem Zeitraster lassen sich die perzeptiven Bewertungen verschiedener Aspekte von Klangfarbe und Dynamik gut erklären, mit einer Varianzaufklärung in der Größenordnung von 50–70%. Eine bessere Vorhersage ist auch durch weitere spektrale Maße ohne Spezifikation der musikalischen Besetzung vermutlich kaum möglich, da etwa die Bewertung einer ›schlanken‹ klangfarblichen Tongebung für einen Kontrabass eine andere Bewertung der spektralen Verteilung erfordert als für eine Piccoloflöte – so wie sich die Suche nach konzeptübergreifenden, für unterschiedliche musikalische Inhalte gültigen Klangfarbenmerkmalen generell als schwierig erwiesen hat.

Die Erklärung von höheren strukturellen Gestaltungsmerkmalen wie Artikulation, Phrasierung und Rhythmisierung gelingt durch die hier eingesetzten Deskriptoren nur mäßig, mit einer Varianzaufklärung in der Größenordnung von 25–45%. Für eine bessere Erklärung insbesondere der Phrasierung und Rhythmisierung wären vermutlich Modelle für den zeitlichen Verlauf von Tempo, Dynamik und Klangfarbe erforderlich, jenseits einer elementaren sensorischen Analyse von Momentanwerten und ihrer statistischen Verteilung. Eine bessere Erklärung der Artikulation kann vermutlich durch eine signalbasierte Erfassung von relativen Tondauern bzw. einer dynamischen Hüllkurve von Einzeltönen erreicht werden.

Evaluative Gesamturteile im Hinblick auf die Stärke des musikalischen Ausdrucks oder die ästhetische Präferenz können auch durch das von den Urteilern selbst erarbeitete semantische Differential von Interpretationsmerkmalen nur zu 67% (musikalischer Ausdruck) bzw. 55% (Gesamteindruck) erklärt werden. Eine positive Bewertung (starker Ausdruck, Gefallen) korreliert in beiden Fällen mit weiträumiger und deutlich ausgeprägter Phrasierung. Für Ausdrucksstärke ist darüber hinaus eine hohe Variationsbreite von Tempo und Klangfarbe maßgeblich, für das Gefallensurteil eine volle Tongebung und eine ausgeprägte Rhythmisierung. In Anbetracht der begrenzten Rückführbarkeit auf perzeptive Einzelattribute ist es jedoch erstaunlich, dass beide Bewertungen auch durch signalbasierte Merkmale zu 59% (musikalischer Ausdruck) bzw. 51% (Gesamteindruck) erklärt werden können. Offenbar sind auch elementare sensorische Momentaneindrücke – das, was man als ›klangliche Oberfläche‹ beschreiben könnte – für die ästhetische Bewertung von Interpretationen keineswegs irrelevant. Über die Bedeutung höherer semantischer Eigenschaften – was in musikalischen Diskursen etwa als die Stimmigkeit, die Konsequenz, die Werktreue oder die Originalität von Interpretationen beschrieben wird – sagt dies freilich wenig aus, denn das Besondere oder das ›Eigentliche‹ einer musikalischen Interpretation kann auch in den letzten 10% an unaufgeklärter Varianz verborgen sein. Andererseits werden vielversprechende Ansätze entwickelt, durch deren Nutzung eine technische Analyse zunehmend nicht nur sensorischen Prozessen, sondern auch Top-Down-Prozessen, mithin dem individuellen Kontext des In-

terpreten und des Hörers, Rechnung tragen könnte: die Einbeziehung komplexerer akustischer Features, die Weiterentwicklung und Implementation algorithmischer Wahrnehmungsmodelle, die Anwendung von Modellen expressiver Gestaltung, die Berücksichtigung von Besonderheiten der Instrumentation und die Erfassung von Maßen der Konventionalität bzw. Originalität von Interpretationen durch Auswertung großer Musikdatenbanken.

Hierdurch bietet sich die Chance, einige bei der Beschreibung von Interpretationen häufig sehr subjektive und ins Mystifizierende tendierende Kategorien zuverlässiger und präziser an das akustische Geschehen rückzubinden als dies bereits auf dem heutigen Stand der Signalverarbeitung möglich ist.[19]

Anhang: Analysierte Einspielungen

Wolfgang Amadeus Mozart, Klaviersonate F-Dur KV 332

Vladimir Horowitz, Classica D'Oro CDO 3304, aufgenommen 1947
Walter Gieseking, EMI 50999 2 65081 2 2, aufgenommen 1953
Lili Kraus, Music & Arts CD-1001(5), aufgenommen 1954
Lili Kraus, Sony Classical CD 62921, aufgenommen 1967/68
Glenn Gould, Sony Classical CD 62921, aufgenommen 1972
Cecile Ousset, Berlin Classics 0093762BC, aufgenommen 1973
András Schiff, Decca 421 110-2, aufgenommen 1980
Friedrich Gulda, Deutsche Grammophon 477 613 0, aufgenommen 1980
Mitsuko Uchida, Philips 412123-2, aufgenommen 1983
Christian Zacharias, EMI Classics 575 041 2, aufgenommen 1984/85
Daniel Barenboim, EMI-Electrola CDS 7473368, aufgenommen 1984
Mieczyslaw Horszowski, Nonesuch 79202, aufgenommen 1988
Maria João Pires, Deutsche Grammophon 477 660 7, aufgenommen 1990
Alfred Brendel, Philips 468 048-2, aufgenommen 2000
Lars Vogt, EMI 0946 3 36080 2 3, aufgenommen 2005
Mikhail Pletnev, Deutsche Grammophon 00289 477 5788, aufgenommen 2005

Ludwig van Beethoven, Streichquartett B-Dur op. 130

Busch Quartett, Sony MPK47687, aufgenommen 1941
Quatuor Vegh , Music & Arts CD-10847, aufgenommen 1952
Hollywood String Quartet, Testament SBT 3082, aufgenommen 1957
Amadeus Quartett, Deutsche Grammophon 4631432, aufgenommen 1962
Quartetto Italiano, Philips 4540622, aufgenommen 1969
Juilliard Quartett, Sony S8K87889, aufgenommen 1970
Yale Quartett, Brilliant Classics 99127, aufgenommen 1971
Lasalle Quartett, Deutsche Grammophon 4537682, aufgenommen 1972
Quatuor Vegh, Valois Auvidis V4400, aufgenommen 1973
Smetana Quartett, Denon COCO-79681, aufgenommen 1982

[19] Die Verfasser danken Herrn Wilko Trebbin für die Durchführung des Hörversuchs und Herrn Fabian Brinkmann für die Mitarbeit an der statistischen Auswertung.

Melos Quartett, Deutsche Grammophon 415676-2, aufgenommen 1985
Guarneri Quartett, Decca 4429402, aufgenommen 1987
Alban Berg Quartett, EMI 5736062, aufgenommen 1989
Lindsay String Quartet, ASV 602, released 1991
Tokyo String Quartet, RCA RD609753, aufgenommen 1990/91
Emerson Quartett, Deutsche Grammophon 474341-2, aufgenommen 1994

Robert Schumann, Fünf Stücke im Volkston op.102

Pablo Casals, Sony Classical SMK 58993, aufgenommen 1952
Mstislav Rostropowitsch, Decca 475 823 9, aufgenommen 1961
Pierre Fournier, Deutsche Grammophon 477 593 9, aufgenommen 1967
André Navarra, Calliope 8361333, aufgenommen 1978
Friedrich-Jürgen Sellheim, Sony Classical SBK 48171, aufgenommen 1978
Raphael Wallfisch, Chandos CHAN 8528, aufgenommen 1986
Yo Yo Ma, Sony Classical SK 92757, aufgenommen 1986
Klaus Storck, Colosseum COL 34.9504, released 1988
Truls Mørk, Simax PSC 1063, aufgenommen 1990
Maria Kliegel, Naxos 8550654, aufgenommen 1991
Jan Vogler, Berlin Classics 0013492BC, aufgenommen 1993
Anner Bylsma, Sony Classical 5173539, aufgenommen 1995
Steven Isserlis, RCA Victor 09026 68800 2, aufgenommen 1997
Mischa Maisky, Deutsche Grammophon 469 524-2, aufgenommen 1997
Daniel Müller-Schott, Orfeo C 617 041 A, aufgenommen 2003
Jean-Guihen Queyras, Alpha Productions AHP 121, aufgenommen 2007