Forschungsberichte aus der
Industriellen Informationstechnik | 34

Markus Schwabe

Analyse und Separation polyphoner Musiksignale

KIT Scientific Publishing

Markus Schwabe

Analyse und Separation polyphoner Musiksignale

Forschungsberichte aus der Industriellen Informationstechnik
Band 34

Institut für Industrielle Informationstechnik
Karlsruher Institut für Technologie
Hrsg. Prof. Dr.-Ing. Michael Heizmann

Eine Übersicht aller bisher in dieser Schriftenreihe erschienenen Bände finden Sie am Ende des Buchs.

Analyse und Separation polyphoner Musiksignale

von
Markus Schwabe

KIT Scientific Publishing

Karlsruher Institut für Technologie
Institut für Industrielle Informationstechnik

Analyse und Separation polyphoner Musiksignale

Zur Erlangung des akademischen Grades eines Doktors der Ingenieurwissenschaften von der KIT-Fakultät für Elektrotechnik und Informationstechnik des Karlsruher Instituts für Technologie (KIT) genehmigte Dissertation

von Markus Schwabe, M.Sc.

Tag der mündlichen Prüfung: 8. März 2024
Hauptreferent: Prof. Dr.-Ing. Michael Heizmann, KIT
Korreferent: Prof. Dr. Christoph Seibert, HfM Karlsruhe

Impressum

Scientific Publishing

Karlsruher Institut für Technologie (KIT)
KIT Scientific Publishing
Straße am Forum 2
D-76131 Karlsruhe

KIT Scientific Publishing is a registered trademark
of Karlsruhe Institute of Technology.
Reprint using the book cover is not allowed.

www.ksp.kit.edu

Print on Demand 2024 – Gedruckt auf FSC-zertifiziertem Papier

ISSN 2190-6629
ISBN 978-3-7315-1365-0
DOI 10.5445/KSP/1000170636

Vorwort

Die vorliegende Arbeit entstand während meiner Zeit als wissenschaftlicher Mitarbeiter am Institut für Industrielle Informationstechnik (IIIT) des Karlsruher Instituts für Technologie (KIT). Dabei haben mich zahlreiche Personen unterstützt, bei denen ich mich bedanken möchte.

Mein erster Dank gilt Prof. Dr.-Ing. Fernando Puente León, der mir durch mein Thema die Verbindung von Hobby und Beruf ermöglichte und mich in den ersten drei Jahren betreute. Ich werde ihn als kompetenten Wissenschaftler und begeisterten Musiker in Erinnerung behalten.

Ganz besonders möchte ich mich bei Prof. Dr.-Ing. Michael Heizmann für die Übernahme der Betreuung während meiner restlichen Zeit am IIIT bedanken. Trotz vieler Verpflichtungen in der für das gesamte Institut schwierigen Zeit hatte er immer ein offenes Ohr für alle meine Anliegen und gab mir wertvolle Impulse für meine Forschungsarbeiten. Des Weiteren bedanke ich mich bei Prof. Dr. Christoph Seibert für die Übernahme des Korreferats und sein Interesse an meiner Arbeit.

Vielen Dank an alle Kolleginnen und Kollegen des IIIT für die stets angenehme Atmosphäre, die große Hilfsbereitschaft sowie die gute und respektvolle Zusammenarbeit. Insbesondere bedanke ich mich bei Johannes Anastasiadis, Manuel Bihler, Muen Jin, Fabian Leven, Daniel Leyer, Lanxiao Li, Sebastian Murgul, Theresa Panther, Johannes Steffens, Erik Tabuchi Barczak, David Uhlig und Hannes Weinreuter für das Korrekturlesen dieser Arbeit. Für die Unterstützung in organisatorischen und technischen Belangen danke ich Manuela Moritz, Patricia Nestl, Dieter Brandt, Marvin Winkler und Stefan Ziegler. Ein weiterer Dank geht an alle Studierenden, die meine Arbeit durch ihre Experimente und viele konstruktive Diskussionen vorangebracht haben.

Darüber hinaus bedanke ich mich bei meinen Eltern und meiner Frau Melinda für ihre Unterstützung und ihren fortwährenden Rückhalt.

Karlsruhe, Mai 2024 Markus Schwabe

Kurzfassung

In dieser Arbeit werden Ansätze zur verbesserten Signalanalyse mehrstimmiger Musikaufnahmen vorgestellt, die auf künstlichen neuronalen Netzen basieren. Diese Ansätze ermöglichen eine objektive Bewertung der Aufnahmequalität von Amateuraufnahmen, eine verbesserte zeitabhängige Detektion aktiver Musikinstrumente sowie eine bessere Separation von Ensemble-Aufnahmen mit unterschiedlichen Instrumenten.

Zur objektiven Qualitätsbewertung unbekannter Musikaufnahmen wird die Schätzung des durchschnittlichen Signal-Rausch-Verhältnisses, des zeitvarianten Signal-Rausch-Verhältnisses kurzer Abschnitte und der Nachhallzeit analysiert. Anschließend wird die Qualitätsbewertung mithilfe einer Beispielanwendung der Klaviertranskription validiert, indem aus den für Aufnahmen unterschiedlicher Qualität geschätzten Parametern die resultierende Transkriptionsgüte abgeleitet wird.

Für die zeitabhängige Instrumentendetektion werden Modelle mit Einspeisung von Zeit-Frequenz-Darstellungen des Musiksignals sowie Modelle mit direkter Einspeisung des Musiksignals miteinander verglichen. Neben der Überlegenheit der Zeit-Frequenz-Darstellungen zeigt sich dabei, dass für die Instrumentendetektion eine hohe Frequenzauflösung der Zeit-Frequenz-Darstellung von Vorteil ist. Darüber hinaus wird eine Steigerung der Detektionsgüte durch die parallele Zusatzeinspeisung verschiedener Phasendarstellungen untersucht.

Im Separationsansatz dieser Arbeit wird die Signaltrennung von unterschiedlichen Musikinstrumenten eines Ensembles verfolgt, die in einem ähnlichen Frequenzspektrum liegen. Dazu werden verschiedene Modellstrukturen untersucht, welche die separierten Signale gleichzeitig oder für jedes Instrument separat schätzen. Eine zusätzliche Einspeisung der zeitabhängigen Instrumentenaktivitäten verbessert die Separationsergebnisse in den meisten Fällen. Der Einfluss der Einspeiseposition wird dabei näher analysiert. Des Weiteren wird die Robustheit der Separationsgüte bei Einspeisung von fehlerhafter Zusatzinformation quantifiziert.

Inhaltsverzeichnis

Vorwort i

Kurzfassung iii

Symbolverzeichnis vii

1 Einleitung 1
- 1.1 Problemstellung 2
- 1.2 Eigener Beitrag 4
- 1.3 Struktur der Arbeit 7

2 Grundlagen 9
- 2.1 Akustik 9
  - 2.1.1 Raumakustik 11
  - 2.1.2 Musiksignale 15
- 2.2 Diskrete Zeit-Frequenz-Darstellungen 17
  - 2.2.1 Kurzzeit-Fourier-Transformation 18
  - 2.2.2 Constant-Q-Transformation 21
- 2.3 Künstliche neuronale Netze 25
  - 2.3.1 Bausteine 26
  - 2.3.2 Architekturen 29
  - 2.3.3 Training 33

3 Bewertung der Qualität polyphoner Musikaufnahmen 35
- 3.1 Qualitätsbewertung von Audiosignalen 36
- 3.2 Schätzung charakteristischer Qualitätsparameter 39
  - 3.2.1 Hintergrundrauschen 41
  - 3.2.2 Kurze Störgeräusche 45
  - 3.2.3 Raumakustik-Parameter 50
- 3.3 Anwendung zur Einordnung der Transkriptionsgüte . . 54

4 Zeitabhängige Instrumentendetektion 65
4.1 Stand der Forschung 66
4.2 Zeit-Frequenz-Darstellungen von Phaseninformation . . 69
4.2.1 Modifizierte Gruppenlaufzeit 71
4.2.2 Produktspektrum 73
4.2.3 Frequenzfehlermatrix 74
4.3 Netzarchitekturen zur Instrumentendetektion 76
4.4 Experimente 82
4.4.1 Implementierungsdetails 82
4.4.2 Vergleich der Netzarchitekturen 85
4.4.3 Einbeziehung von Phaseninformation 95

5 Separation polyphoner Ensemble-Aufnahmen 99
5.1 Quellentrennung von Musiksignalen 100
5.1.1 KNN-basierte Separationsmodelle 102
5.1.2 Separation mit Einspeisung von Vorwissen 105
5.2 Separationsmodell 107
5.3 Datensatz 112
5.4 Multi-Task-Ansatz zur Separation 117
5.5 Separationsergebnisse mit simulierter Zusatzinformation 120
5.5.1 Gemeinsames Separationsmodell 121
5.5.2 Unabhängige Separationsmodelle 129
5.6 Auswirkungen realer Instrumentendetektions-
ergebnisse als Zusatzinformation 138

6 Fazit 145
6.1 Zusammenfassung 145
6.2 Ausblick 147

**A Detaillierte Ergebnisse zur zeitabhängigen
Instrumentendetektion 151**

B Detaillierte Ergebnisse der Separationsmodelle 155

Literaturverzeichnis 167
Eigene Veröffentlichungen 182
Betreute studentische Arbeiten 183

Symbolverzeichnis

Allgemeine Abkürzungen

Abkürzung	Bedeutung
BCE	binäre Kreuzentropie
BSS	blinde Quellentrennung
bzw.	beziehungsweise
CE	Kreuzentropie
CNN	neuronale Faltungsnetze
CQT	Constant-Q-Transformation
d. h.	das heißt
DL	Deep Learning
EDT	Anfangsnachhallzeit
FFT	schnelle Fourier-Transformation
FT	Fourier-Transformation
GLU	Gated Linear Unit
GRU	Gated Recurrent Unit
GD	Gruppenlaufzeit
ID_ZF_Res	Modell zur Instrumentendetektion mit Einspeisung von Zeit-Frequenz-Darstellungen
ID_ZF_Deep	tiefes Modell zur Instrumentendetektion mit Einspeisung von Zeit-Frequenz-Darstellungen
ID_t_Res	Modell zur Instrumentendetektion mit Einspeisung des Zeitsignals
KI	künstliche Intelligenz
KNN	künstliche neuronale Netze
LSTM	Long Short-Term Memory
MAE	mittlerer absoluter Fehler
MIDI	Musical Instrument Digital Interface

Abkürzung	Bedeutung
MIR	Music Information Retrieval
ML	maschinelles Lernen
ModGD	modifizierte Gruppenlaufzeit
IF	Momentanfrequenz
MSE	mittlerer quadratischer Fehler
MT	Multi-Task
NSGT	nichtstationäre Gabor-Transformation
PS	Produktspektrum
ReLU	Rectified Linear Unit
RIR	Raumimpulsantwort
RNN	rekurrente neuronale Netze
SGD	stochastischer Gradientenabstieg
SI-SAR	Scale-Invariant Source-to-Artifacts Ratio
SI-SDR	Scale-Invariant Source-to-Distortion Ratio
SI-SIR	Scale-Invariant Source-to-Interferences Ratio
SNR	Signal-Rausch-Verhältnis
STD	Standardabweichung
STFT	Kurzzeit-Fourier-Transformation
z. B.	zum Beispiel

Symbole

Lateinische Buchstaben

Symbol	Bedeutung
a	Dämpfung
A	Fläche
b	Anzahl Frequenzbins pro Oktave
B	Inharmonizitätsfaktor
c_0	Schallgeschwindigkeit
d	Dimension
e	Signalanteil
f	Frequenz
f_A	Abtastfrequenz

Symbol	Bedeutung
F	F-Maß
g	Raumimpulsantwort
h	Zeitverschiebung des Analysefensters
I_{Instr}	Anzahl betrachteter Instrumente
J	Kostenfunktion
k	Frequenzindex der Zeit-Frequenz-Darstellung
K	Gesamtanzahl der Frequenzbins
m	Zeitindex der Zeit-Frequenz-Darstellung
M	Gesamtanzahl der Zeitbins
n	Zeit (diskret)
N	Signallänge
N_w	Breite des Fensters w
p	Schalldruck
P	Produktspektrum
Q	Verhältnis der Mittenfrequenz zur Frequenzauflösung
r	Integrierte Impulsantwort
s	Quellensignal
t	Zeit (kontinuierlich)
t_{EDT}	Anfangsnachhallzeit
t_{RT}	Nachhallzeit
V	Raumvolumen
w	Fensterfunktion
x	Sensorsignal
X	Transformierte des Sensorsignals x
y	Ausgangssignal
z	Zustand

Griechische Buchstaben

Symbol	Bedeutung
α	trainierbares Gewicht eines künstlichen Neurons
β	Bias eines künstlichen Neurons
χ	Gewichtungsparameter von Gütefunktionen
λ	Wellenlänge

Symbol	Bedeutung
ν	Lernrate
ω	Kreisfrequenz
Ψ	Frequenzfehlermatrix
σ	Aktivierungsfunktion
τ	Laufzeit
θ	Phase der Transformierten X

Hochgestellte Indizes

Index	Bedeutung
$(\bullet)^{(V)}$	Anzahl V der Merkmalskarten eines Tensors
$(\bullet)^{*}$	konjugiert komplex

Tiefgestellte Indizes

Index	Bedeutung
$(\bullet)_{\mathrm{CS}}$	cepstral geglättetes Element
$(\bullet)_{\mathrm{Im}}$	Imaginärteil
$(\bullet)_{\mathrm{Re}}$	Realteil
$(\bullet)_{ji}$	Verbindung oder Parameter von Element i zu Element j
$(\bullet)_{n}$	Element zum Zeitpunkt n

Mathematische Operatoren

Operator	Bedeutung
∂	partielle Ableitung
$\mathcal{F}$	Fourier-Transformation

1 Einleitung

Musik ist schon seit Jahrhunderten ein essenzieller Bestandteil menschlicher Kultur. Heutzutage begleitet sie viele Menschen im täglichen Leben, sowohl beim bewussten Konsum von Musik als auch im Hintergrund, wie z. B. beim Einkaufen oder in der Gastronomie. Die hohe Verbreitung und Verfügbarkeit von Musik wurde im 20. Jahrhundert durch das Radio, die Schallplatte sowie die Entwicklung von digitalen Aufnahme- und Abspieltechniken wie der CD oder dem mp3-Player ermöglicht. In den letzten Jahren hat sich die Verfügbarkeit durch Smartphone und Internetstreaming weiter gesteigert. Darüber hinaus können eigene Musikaufnahmen ohne großen Aufwand, beispielsweise mit dem Smartphone, erstellt und über das Internet einer breiten Masse zur Verfügung gestellt werden. Um die zunehmende Menge an verfügbaren Musikaufnahmen sinnvoll zu strukturieren, sind Eigenschaften wie Genre, Besetzung oder Stimmung der Aufnahme hilfreich. Basierend auf diesen Merkmalen schlagen Suchalgorithmen dem Nutzer ähnliche Musikstücke vor und erstellen automatisiert Wiedergabelisten zu definierten Themen.

Sind die Eigenschaften der Musikaufnahmen nicht bekannt, können sie entweder durch Personen definiert oder mit relativ geringem Aufwand durch spezielle Algorithmen aus den digitalen Musiksignalen geschätzt werden. Solche Algorithmen sind Teil des Signalverarbeitungsfeldes *Music Information Retrieval* (MIR), welches die Extraktion der im Musiksignal enthaltenen Information umfasst. Neben typischen Klassifikationsaufgaben wie der Erkennung des Genres [116], der Besetzung [40] oder des Komponisten [142] verfolgen mehrere MIR-Aufgaben auch die Extraktion von zeitabhängigen Signalverläufen. Dazu gehören unter anderem die Transkription gespielter Noten [44], die Detektion von Akkordfolgen [138] oder die Separation der Musikaufnahme in einzelne Signalkomponenten [18, 132]. Die resultierenden Signale können teilweise auch als Startpunkt für weitere MIR-Aufgaben oder Algorithmen dienen. So erleichtert die Separation einzelner Instrumentensignale unter anderem

die Transkription dieser Instrumente oder die Neuabmischung von Audioaufnahmen wie z. B. durch ein automatisches DJ-Programm [140].

Insgesamt kann eine digitale Musikaufnahme sehr schnell und vergleichsweise einfach nachbearbeitet werden, wenn sie über verschiedene Spuren verfügt. Wenn das nicht der Fall ist und nur eine gemeinsame Aufnahmespur vorliegt, erweitert eine Separation dieses Musiksignals in mehrere Quellensignale den Umfang an möglichen Nachbearbeitungen erheblich. Musikalische Quellen können dabei entweder einzelne Musikinstrumente oder auch eine Gruppe von Instrumenten sein, wie z. B. alle Begleitstimmen, die keine Melodie spielen. Ein Melodieinstrument, häufig gegenüber den anderen Komponenten hervorgehoben, ist in fast allen MIR-Aufgaben einfacher zu behandeln als mehrere Quellen mit gleichzeitig gespielten Melodielinien. Diese Signale mit mehreren parallelen Melodien, egal ob von gleichen oder verschiedenen Instrumenten gespielt, werden als polyphone Musiksignale bezeichnet und in der vorliegenden Arbeit untersucht. Darüber hinaus wird eine breite Anwendbarkeit der entwickelten Ansätze verfolgt, sodass ausschließlich monaurale Musiksignale analysiert werden. Sie besitzen nur einen Aufnahmekanal und können daher mit nur einem Mikrofon, beispielsweise über das Smartphone, aufgenommen werden.

In dieser Arbeit werden mit der Bewertung der Aufnahmequalität von Amateuraufnahmen, der zeitabhängigen Instrumentendetektion und der Separation von Ensemble-Aufnahmen drei wichtige Themengebiete der Analyse von polyphonen Musiksignalen untersucht.

1.1 Problemstellung

Die Qualität von Musikaufnahmen kann die Ergebnisse der angewandten MIR-Algorithmen teilweise stark beeinflussen. Da gerade im Falle von Amateuraufnahmen häufig schwankende Aufnahmequalitäten auftreten, wird im ersten Teil dieser Arbeit eine objektive Bewertung der Aufnahmequalität vorgestellt. Sie ermöglicht die Abschätzung des Einflusses der Qualität auf die Signalverarbeitung der nachfolgenden MIR-Aufgabe. Für die objektive Qualitätsbeurteilung spielen hauptsächlich Störeffekte im Signal eine Rolle, wohingegen die menschliche Wahrnehmung nicht berücksichtigt wird. Als Hauptursachen solcher Störeffekte wurden mo-

notones Hintergrundrauschen, kurze Störgeräusche mit variierender Charakteristik und raumakustische Einflüsse wie der Nachhall identifiziert. Diese sollen zur Bewertung der Qualität anhand von charakteristischen Parametern separat geschätzt werden. Die getrennte Schätzung ermöglicht eine gezielte Änderung von Aufnahmeparametern, um Musiksignale mit bestmöglicher Qualität aufnehmen zu können. Das ist beispielsweise für Amateuraufnahmen mit dem Smartphone interessant, da das Aufnahmeszenario meist relativ einfach variiert werden kann. Alle Schätzungen der definierten Qualitätsparameter werden anhand von mit entsprechenden Störeffekten überlagerten Klaviersignalen getrennt evaluiert. Als Beispiel für den Einfluss auf eine nachfolgende MIR-Anwendung wird die Transkriptionsgüte von polyphonen Klaviersignalen bei unterschiedlicher Aufnahmequalität analysiert und mit den geschätzten Qualitätsparametern in Beziehung gesetzt.

Polyphone Musiksignale enthalten oft unterschiedliche Musikinstrumente, welche parallele Melodien spielen. Diese Instrumente sind häufig nicht vorab bekannt, was eine Einordnung der Musikaufnahme anhand ihrer Besetzung unmöglich macht. Darüber hinaus ist die Information über die genauen Zeitabschnitte, in denen die Instrumente während der Aufnahme aktiv sind, nur in Ausnahmefällen gegeben. Vielen MIR-Algorithmen würde dieser Aktivitätsverlauf aber als zusätzliche Information helfen, da sie sich auf die relevanten Abschnitte fokussieren könnten. Aus diesem Grund wird im zweiten Teil der Arbeit eine zeitabhängige Detektion der während der Aufnahme aktiven Musikinstrumente entwickelt. Dabei soll der Status jedes Instruments während eines Zeitschrittes entweder aktiv oder nicht aktiv sein, was einer typischen Klassifikationsaufgabe entspricht. Dazu werden verschiedene Detektionsmodelle auf Basis von unterschiedlichen Zeit-Frequenz-Darstellungen sowie direkt eingespeisten Zeitsignalen verglichen. Die Zeit-Frequenz-Darstellungen enthalten dabei nur die Betragswerte der zugehörigen Signaltransformation, weshalb die Auswirkung von zusätzlichen Eingangsdarstellungen der Phasenwerte ebenfalls untersucht wird.

Um einzelne Stimmen in monauralen Musikaufnahmen mit mehreren Instrumententypen auch nachträglich noch bearbeiten zu können, ist eine digitale Separation der Instrumentenstimmen notwendig. Ein Separationsansatz mit direkter Schätzung der Zeitsignale wird im dritten

Teil dieser Arbeit entwickelt. Der Fokus des Separationsansatzes liegt hier auf der Trennung von unterschiedlichen Instrumenten in Ensemble-Aufnahmen der Kammermusik, weshalb die in der Literatur üblichen Komponenten Gesang oder Schlagzeug nicht berücksichtigt werden. Dadurch sind die charakteristischen Eigenschaften der zu trennenden Komponenten deutlich ähnlicher, was die Quellenseparation herausfordernder macht. Da nur die Separation unterschiedlicher Instrumententypen umgesetzt wird, werden Signale von gleichen Instrumenten immer als eine Quelle aufgefasst. Im Rahmen dieser Arbeit werden verschiedene Modellarchitekturen anhand der geschätzten Instrumentensignale verglichen. Darüber hinaus wird die Erweiterung des Separationsansatzes durch die Integration von Zusatzinformation über aktive Instrumente vorgeschlagen. Diese erfordert kein zusätzliches Vorwissen über das Musiksignal, wenn die Instrumentenaktivität wie im zweiten Teil der Arbeit aus dem Signal geschätzt wird. Eine Schätzung enthält in der Regel aber Fehler, weshalb die Robustheit der Separation gegenüber zufälligen Fehlern in der Zusatzinformation untersucht und mithilfe von realen Schätzungen validiert wird.

Aufgrund der Betrachtung von monauralen Musiksignalen kann in keinem der drei vorgestellten Ansätze eine räumliche Information der Klangquellen ausgenutzt werden. Diese Einschränkung stellt im Vergleich zu mehrkanaligen Aufnahmen einen Nachteil dar, ermöglicht aber den universellen Einsatz der Ansätze auf alle digitalen Musiksignale.

1.2 Eigener Beitrag

Alle drei in dieser Arbeit untersuchten Gebiete zur Analyse polyphoner Musiksignale werden durch die entwickelten Ansätze vorangebracht. Für die Bewertung der Aufnahmequalität sind die wichtigsten Beiträge dieser Arbeit:

- Es werden Qualitätsparameter zur objektiven Bewertung von Musikaufnahmen hinsichtlich Hintergrundrauschen, kurzen Störgeräuschen und Nachhall definiert, die eine störungsbezogene Abschätzung von zu erwartenden MIR-Ergebnissen ermöglichen.

- Die zuverlässige Schätzung des Signal-Rausch-Verhältnis von unbekannten Musikaufnahmen wird für stationäres Hintergrundrauschen und zeitvariante Störcharakteristiken als erster Ansatz mithilfe eines künstlichen neuronalen Netzes (KNN) umgesetzt, das darüber hinaus nur wenige Parameter besitzt.
- Zur Schätzung von Nachhall- oder Anfangsnachhallzeit wird ebenfalls jeweils ein KNN mit wenigen Parametern vorgestellt, welches nach Kenntnis des Autors der bisher einzige Ansatz für unbekannte Musikaufnahmen ist.
- Die geschätzten Qualitätsparameter werden erfolgreich auf die Beispielanwendung der Klaviertranskription übertragen, indem sie zur Einschätzung der resultierenden Transkriptionsgüte bei Klavieraufnahmen unterschiedlicher Qualität eingesetzt werden.

Im Falle der zeitabhängigen Instrumentendetektion werden die Untersuchungen dieser Arbeit in Anlehnung an das Vorgehen von Hung und Yang [53] durchgeführt. Dabei sind die wichtigsten Beiträge:

- Die zeitabhängige Schätzung der Instrumentenaktivität berücksichtigt Nachbarinformationen der Signaldarstellung mithilfe von zweidimensionalen Faltungsschichten in den *Residual*-Modulen.
- Ein ausführlicher Vergleich von zwei unterschiedlich tiefen Modellarchitekturen mit Einspeisung von Zeit-Frequenz-Darstellungen sowie einem Modell mit direkter Einspeisung des Zeitsignals zeigt die Überlegenheit der Signalvorverarbeitung für diese Aufgabe.
- Für die Zeit-Frequenz-Darstellungen aus Kurzzeit-Fourier-Transformation (STFT) und Constant-Q-Transformation werden unterschiedliche Eingangsdimensionen hinsichtlich der erzielbaren Instrumentendetektion untersucht. Dabei wird die Wichtigkeit einer hohen Frequenzauflösung deutlich.
- Um die normalerweise vernachlässigte Phaseninformation der Signaltransformation in die Schätzung zu integrieren, wird eine zusätzliche Phasendarstellung parallel zur STFT in das KNN eingespeist. Im Falle einer geringen Frequenzauflösung verbessern alle drei untersuchten Phasendarstellungen die Ergebnisse.

- Die besten Detektionsmodelle dieser Arbeit schätzen eine um 15 % höhere Zeitauflösung als der Vergleichsansatz der Literatur [53] und erzielen trotzdem bessere Ergebnisse.

Der Separationsansatz basiert auf dem Literaturmodell Demucs [18], das ohne Signaltransformation auskommt und die Quellensignale direkt schätzt. Das in dieser Arbeit entwickelte System zeichnet sich hauptsächlich durch folgende Beiträge aus:

- Anstatt der in der Literatur üblichen Separation von Gesang, Bass, Schlagzeug und Rest wird ein Ansatz zur Trennung von Ensemble-Aufnahmen unterschiedlicher Musikinstrumente eingeführt, die in einem ähnlichen Frequenzspektrum liegen.
- Mit einem gemeinsamen Gesamtmodell, einem Multi-Task-Ansatz und einem Separationssystem mit unabhängigen Einzelmodellen werden drei verschiedene Modellstrukturen untersucht und anhand ihrer Separationsergebnisse verglichen.
- Durch die zusätzliche Einspeisung der zeitabhängigen Aktivitätsinformation aller zu trennenden Instrumente werden die Separationsergebnisse verbessert. Die Integration dieser Zusatzinformation kann theoretisch vor jedem Block des Encoders erfolgen, weshalb die Ergebnisse aller Einspeiseorte analysiert werden.
- Bei Einspeisung von fehlerhafter Zusatzinformation reduziert sich die Separationsqualität je nach Einspeiseort unterschiedlich stark. Deshalb wird die Robustheit der Separationsergebnisse in der Simulation gegenüber zufällig invertierten Werten analysiert.
- Die auf simulierten Fehlern basierende Robustheitsanalyse wird anhand von echten Schätzungen der Instrumentenaktivitäten verifiziert, die mithilfe der im zweiten Teil der Arbeit entwickelten Instrumentendetektion generiert werden.

1.3 Struktur der Arbeit

Da die vorliegende Arbeit mit der Bewertung der Aufnahmequalität, der zeitabhängigen Instrumentendetektion und der Separation polyphoner Musikaufnahmen drei nicht unmittelbar zusammenhängende Teilgebiete der MIR untersucht, wird jedes Gebiet in einem separaten Kapitel behandelt. Zuvor werden in Kapitel 2 die Grundlagen erläutert, die in allen Teilgebieten benötigt werden. Nach einer Einführung in die Akustik, in der die Raumakustik und Musiksignale näher beleuchtet werden, folgt die Vorstellung der in dieser Arbeit verwendeten Signaltransformationen der Kurzzeit-Fourier- und der Constant-Q-Transformation. Anschließend wird auf die wichtigsten Bausteine und Architekturen von künstlichen neuronalen Netzen sowie ihr Training eingegangen.

Kapitel 3 beinhaltet die Qualitätsbewertung von polyphonen Musikaufnahmen. Zunächst wird ein Überblick über bestehende Verfahren zur Qualitätsbewertung von Audiosignalen gegeben. Danach wird die unabhängige Schätzung der Qualitätsparameter für die drei definierten Einflussfaktoren Hintergrundrauschen, kurze Störgeräusche und Nachhall in je einem Abschnitt vorgestellt und evaluiert. Diese Parameterschätzungen werden dann zur Abschätzung der Transkriptionsgüte einer Klaviertranskription verwendet und anhand dieser validiert.

Das Kapitel 4 zur zeitabhängigen Instrumentendetektion beginnt mit einem Abschnitt über den Stand der Forschung zu diesem Thema. Im Anschluss werden spezielle Zeit-Frequenz-Darstellungen beschrieben, die neben der Betrags- auch die Phaseninformation der zugrunde liegenden Signaltransformation miteinbeziehen. Danach werden die drei Netzarchitekturen vorgestellt, die in dieser Arbeit zur Schätzung der Instrumentenaktivität vorgeschlagen und untersucht werden. Diese kommen in den nachfolgenden Experimenten zum Einsatz, welche sowohl einen Vergleich der unterschiedlichen Netzarchitekturen als auch eine Analyse der zusätzlichen Einspeisung aller vorab beschriebenen Phasendarstellungen enthalten.

Als drittes MIR-Themengebiet wird in Kapitel 5 die Separation polyphoner Ensemble-Aufnahmen behandelt. Der Überblick über bestehende Arbeiten ist dabei in KNN-basierte Separationsansätze und Verfahren mit Einspeisung von Vorwissen unterteilt. Danach wird das in dieser Arbeit weiterentwickelte Separationsmodell sowie der in den Experimenten

verwendete Datensatz vorgestellt. Mit dem Multi-Task-Ansatz wird zudem eine spezielle Umsetzung des Separationsmodells untersucht. Seine Separationsergebnisse werden mit denen der zuvor vorgestellten Umsetzungen eines gemeinsamen Separationsmodells aller zu trennenden Instrumente sowie eines Separationssystems bestehend aus unabhängigen Einzelmodellen verglichen. Anschließend wird die Verbesserung der Separation dieser Separationsansätze durch die zusätzliche Einspeisung von Informationen über die zeitabhängigen Instrumentenaktivitäten untersucht. Dabei werden sowohl die Auswirkungen von simulierten Zusatzinformationen mit zufälligen Fehlern als auch die Auswirkungen von realen Detektionsergebnissen analysiert.

Die Ergebnisse der gesamten Arbeit werden schließlich in Kapitel 6 zusammengefasst. Darüber hinaus wird ein Ausblick auf mögliche weiterführende Forschungsarbeiten gegeben.

2 Grundlagen

In diesem Kapitel werden allgemeine Grundlagen erläutert, die zum Verständnis der Arbeit wichtig sind. Zunächst gibt Abschnitt 2.1 eine kurze Einführung in die Akustik, welche die Basis für Musikwahrnehmung darstellt. Danach werden in Abschnitt 2.2 die beiden wichtigsten in dieser Arbeit verwendeten Zeit-Frequenz-Darstellungen vorgestellt. Abschließend geht Abschnitt 2.3 auf die Grundlagen künstlicher neuronaler Netze und ihrer Architekturen ein.

2.1 Akustik

Die Lehre vom Schall wird als Akustik bezeichnet. Sie umfasst sowohl Eigenschaften als auch Generierung, Übertragung, Messung und Wahrnehmung von Schall [74]. Eine der wichtigsten Eigenschaften des Schalls ist seine Frequenz, mit welcher die von der Schallwelle angeregten Teilchen in den übertragenden Gasen, Flüssigkeiten oder Festkörpern schwingen. Für die Übertragung von Musiksignalen ist nur die Ausbreitung in Luft relevant, weshalb sich die Ausführungen im Folgenden nur auf den Luftschall beziehen. Speziell in der Audiotechnik kann dies weiter auf den hörbaren Luftschall zwischen 16 Hz und 20 kHz eingegrenzt werden [149], da Infraschall mit tieferen Frequenzen als 16 Hz und Ultra- bzw. Hyperschall mit Frequenzen über 20 kHz vom menschlichen Ohr in der Regel nicht wahrgenommen werden können. Einen detaillierten Überblick über die Akustik in allen Übertragungsmedien geben Lerch et al. [74] oder Möser [96].

Luftschall breitet sich als longitudinale Kompressionswelle aus, was in einer ortsabhängigen Schwankung des Luftdrucks und der Dichte resultiert. Dabei schwingen von der Welle angeregte Luftmoleküle entlang der Ausbreitungsrichtung und regen ihre benachbarten Moleküle über elastische Stöße an, sodass eine Kettenreaktion entsteht [149]. Mathema-

tisch kann diese Ausbreitung mithilfe der skalaren Wellengleichung für den Schalldruck p

$$\frac{1}{c_0^2}\frac{\partial^2 p}{\partial t^2} - \Delta p = 0 \tag{2.1}$$

oder äquivalenten Wellengleichungen für z. B. die Dichteänderung beschrieben werden [74]. Dabei steht $\Delta = \operatorname{div}\operatorname{grad}$ für den Laplace-Operator. Die mithilfe von Gleichung (2.1) definierte Welle breitet sich mit der Phasengeschwindigkeit c_0 aus. Diese hängt im Falle des Luftschalls von der Luftzusammensetzung, der Luftfeuchtigkeit und hauptsächlich von der Temperatur ab. Unter normalen atmosphärischen Bedingungen, einer Luftfeuchte von 50 % und einer Raumtemperatur von 20 °C beträgt die Geschwindigkeit etwa 343,6 m s^{-1} und steigt näherungsweise linear um 0,6 m s^{-1} pro 1 K Raumtemperaturerhöhung [149].

Soll nicht die detaillierte Druckverteilung im Raum, sondern rein die Schallausbreitung beschrieben werden, ist eine Betrachtung der bisher dargestellten Wellenakustik meist überflüssig. Hier bietet die geometrische Akustik eine deutlich übersichtlichere Darstellung, da sie die Ausbreitung in Schallstrahlen modelliert, die an Objekten nach dem Reflexionsgesetz vollständig oder teilweise reflektiert werden. Diese Beschreibung ist immer dann möglich, wenn die Schallwellenlängen klein gegenüber den mit dem Schall wechselwirkenden Objekten sind [74]. Hörbarer Schall besitzt nach

$$c_0 = \lambda \cdot f \tag{2.2}$$

Wellenlängen im Bereich zwischen 21,5 m und 17 mm, sodass die geometrische Akustik in den meisten Fällen gilt, aber bei sehr tiefen Frequenzen häufig nicht mehr verwendet werden kann. Zur Beschreibung der Akustik in begrenzten Räumen, auf die in Abschnitt 2.1.1 näher eingegangen wird, stellt sie dennoch häufig die Grundlage dar.

Um den Luftschall zu messen und daraus ein digitales Signal zu gewinnen, werden Mikrofone eingesetzt, welche die akustische Energie der Welle in elektrische Energie wandeln. Das dabei eingesetzte Wandlerprinzip hängt vom Mikrofontyp ab. In der Praxis werden am häufigsten Kondensatormikrofone und elektrodynamische Mikrofone eingesetzt [96], wobei es darüber hinaus zahlreiche weitere Umsetzungen wie z. B. piezoelektrische oder optische Mikrofone gibt. Neben dem Wandlerprin-

zip können Mikrofone auch anhand ihrer Richtcharakteristik oder ihrer Bauform klassifiziert werden. In dieser Arbeit wird ausschließlich das resultierende Musiksignal analysiert, weshalb für eine tiefergehende Beschreibung von Mikrofonen auf die Literatur [74, 96, 149] verwiesen wird. Eigenschaften von Musiksignalen werden in Abschnitt 2.1.2 vorgestellt.

2.1.1 Raumakustik

Im idealen Fall des Freifelds emittiert eine Schallquelle Wellen, die sich radial ausbreiten und sich somit immer weiter von der Quelle entfernen. Dabei kann die Schallabstrahlung kugelförmig, d. h. in alle Richtungen gleich, oder mit einer Richtungsabhängigkeit erfolgen, die in der spezifischen Richtcharakteristik der Quelle beschrieben wird. Ein im Feld platzierter Sensor j nimmt dann mit

$$x_j(t) = a_{ji}\, s_i\Big(t - t_{ji}\Big) \tag{2.3}$$

nur den um a_{ji} gedämpften und t_{ji} zeitverzögerten Direktschall des Quellsignals $s_i(t)$ auf, das in seine Richtung emittiert wurde.

Die Aufnahme von Musiksignalen erfolgt aber in der Regel in begrenzten Räumen, deren Schallausbreitung von vielfachen Reflexionen des emittierten Schalls an Wänden und Objekten gekennzeichnet ist. Dadurch überlagern sich Direktschall und reflektierte Schallwellen, die eine beliebige Anzahl an Reflexionen erfahren können und im betrachteten Punkt zusammentreffen. Unter Annahme der geometrischen Akustik lässt sich das resultierende Signal an einem Sensor j als Mehrwegeausbreitung von L Ausbreitungspfaden beschreiben, woraus sich die Summe

$$x_j(t) = \sum_{l=1}^{L} a_{ji}^{l}\, s_i\Big(t - t_{ji}^{l}\Big) \tag{2.4}$$

ergibt. Jeder Pfad l hat dabei aufgrund der unterschiedlichen Ausbreitungswege seine individuelle Dämpfung a_{ji}^{l} und Zeitverzögerung t_{ji}^{l}. Vereinfachend wird die Dämpfung hierbei als frequenzunabhängig angenommen, obwohl sich die Frequenzcharakteristik eines Signals durch Reflexionen oft leicht verändert [71].

Für den Fall sehr kleiner Dämpfungen a_{ji} wird die Welle der Schallquelle sehr oft reflektiert, sodass es am beobachteten Punkt so scheint,

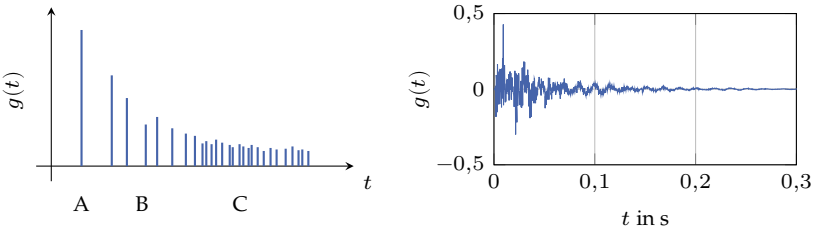
als würden Wellen aus allen Richtungen mit nahezu gleicher Intensität eintreffen. Dieser Sonderfall erzeugt ein sogenanntes „diffuses Schallfeld“ im Raum, das sowohl einen gleichverteilten örtlichen Pegel als auch eine gleichverteilte Einfallsrichtung besitzt [96].

Eine allgemeinere und oft einfachere Beschreibung der Raumakustik bietet die Raumimpulsantwort (englisch *room impulse response*, RIR). Sie basiert auf der Interpretation des Raumes als lineares und zeitinvariantes akustisches Übertragungssystem, was in guter Näherung erfüllt ist [149]. Dadurch lässt sich das Signal am Sensor j mithilfe der Faltung

$$x_j(t) = g_{ji}(t) * s_i(t) \tag{2.5}$$

berechnen. Die RIR $g_{ji}(t)$ beschreibt alle Signaländerungen, die durch Reflexion und Ausbreitung der emittierten Schallwelle zwischen Quelle i und Sensor j hervorgerufen werden. Aus ihr lassen sich nahezu alle raumakustischen Kriterien und Parameter ableiten [71]. Allerdings gilt diese vollständige Beschreibung des Raumes nur für die jeweils betrachtete, feste Anordnung von Quelle und Sensor im Raum.

Die Raumimpulsantwort kann entweder im betreffenden Raum gemessen oder anhand von Computersimulationen des Raumes berechnet werden. Messtechnisch lässt sich die RIR mithilfe unterschiedlicher Verfahren bestimmen, welche die Antwort auf ein definiertes Anregungssignal aufzeichnen und daraus die Impulsantwort errechnen. Mögliche Anregungssignale sind ein lauter Knall, der einen Dirac-Impuls annähert, pseudozufällige Binärfolgen wie die *Maximum Length Sequence* oder Sinussignale mit langsamer Gleitfrequenz, das sogenannte *Sine Sweep*. Diese und weitere Verfahren wurden von Stan et al. [129] zusammengefasst und miteinander verglichen. Dagegen verfolgt die synthetische RIR-Generierung im Wesentlichen zwei Hauptansätze, bei denen die Raumbedingungen als Randbedingungen einfließen. Ein Ansatz basiert auf der numerischen Lösung der Wellengleichungen, was sehr genau, aber auch sehr rechenaufwendig ist. Beispiele hierfür sind die Finite-Elemente-Methode [74] oder die Randelemente-Methode [67]. Der zweite Ansatz ist strahlenbasiert und beruht auf der geometrischen Akustik. Dabei wird die Schallausbreitung aller möglichen Pfade z. B. durch die Mehrwegeausbreitung aus Gleichung (2.4) simuliert und aufsummiert. Bekannte Verfahren sind das Raytracing oder Beamtracing, die im Überblick von Savioja und Svensson [115] näher erläutert werden.



(a) Schema einer RIR. **(b)** Reale RIR.

Abbildung 2.1 Schematische Struktur einer Raumimpulsantwort und reale RIR-Aufnahme eines Hotelzimmers aus dem RIR-Datensatz von Szöke et al. [135].

Jede RIR kann schematisch in drei Bereiche unterteilt werden, in denen jeweils unterschiedliche Arten des Schalls präsent sind. Der Direktschall (A) erfährt keine Reflexion, weshalb die Schallwelle den Sensor am schnellsten und mit der geringsten Dämpfung erreicht. Schallanteile, die nur wenige Reflexionen erfahren, werden als frühe Reflexionen (B) bezeichnet, da sie nur kurz nach dem Direktschall am Empfänger ankommen. Danach schließt sich der diffuse Schallanteil an, der als Nachhall (C) wahrgenommen wird [71]. Die Bereiche sind in der schematischen Struktur in Abbildung 2.1(a) markiert. Zum Vergleich ist in Abbildung 2.1(b) eine reale Aufnahme der RIR eines Hotelzimmers aus dem in Abschnitt 3.2.3 verwendeten RIR-Datensatz [135] dargestellt.

In den abgebildeten Raumimpulsantworten wird deutlich, dass die Signalenergie bei plötzlicher Abschaltung der Schallquelle näherungsweise exponentiell mit der Zeit abklingt. Der damit verbundene Abfall der Schallenergiedichte wird häufig als Nachhall bezeichnet und stellt eines der wichtigsten raumakustischen Charakteristika dar. Als zugehörigen Parameter hat Sabine [113] hierfür schon um 1900 die Nachhallzeit t_{RT} definiert. Sie gibt an, nach welcher Zeit die Signalenergie um 60 dB abgesunken ist. Sabine bestimmte empirisch die Berechnungsformel

$$t_{\mathrm{RT}} = 0{,}163\,\frac{V}{A_{\mathrm{ges}} + 4\,a_{\mathrm{Luft}}\,V}\,, \tag{2.6}$$

die abhängig vom Raumvolumen V, der äquivalenten Schallabsorbtionsfläche A_{ges} und der Luftdämpfung a_{Luft} ist. Eine genauere Berechnung

ermöglicht die Methode der integrierten Impulsantwort [119], bei der zunächst das Integral

$$r(t) = \int_t^\infty g^2(t')\mathrm{d}t' = \int_0^\infty g^2(t')\mathrm{d}t' - \int_0^t g^2(t')\mathrm{d}t' \tag{2.7}$$

über die quadrierte Raumimpulsantwort $g(t)$ gebildet wird. Das Ergebnis $r(t)$ enthält die zum Zeitpunkt t noch im Raum enthaltene Schallenergie, woraus die gesuchte Zeit für 60 dB Energieabfall leicht mithilfe von

$$r_{\mathrm{dB}}(t) = 10 \cdot \log\left(\frac{r(t)}{r(0)}\right) \tag{2.8}$$

berechnet werden kann. Im Allgemeinen ist die Nachhallzeit frequenzabhängig und muss für unterschiedliche Frequenzbereiche separat bestimmt werden. Sie kann aber auch als Einzahlwert angegeben werden, wofür meist der Mittelwert der Nachhallzeiten für 500 Hz und 1000 Hz berechnet wird [149].

Da für einen Energieabfall von 60 dB Rauschen gegen Ende des Nachhalls eine immer stärkere Rolle spielt und folglich eine genaue Messung erschwert, wird häufig nur der Abfall über einen kleineren Bereich gemessen und dieser Wert dann extrapoliert. Dazu eignen sich laut DIN EN ISO 3382 die Energieintervalle zwischen –5 dB und –15 dB, –25 dB oder –35 dB, woraus die Nachhallzeiten $t_{\mathrm{RT}_{10}}$, $t_{\mathrm{RT}_{20}}$ bzw. $t_{\mathrm{RT}_{30}}$ resultieren. Alternativ kann bei sehr kleinen Lautstärken auch die Anfangsnachhallzeit (englisch *early decay time*, EDT) verwendet werden, welche die Zeit für ein Abklingen der Signalenergie auf –10 dB repräsentiert. Im Vergleich zur Nachhallzeit ist die EDT deutlich abhängiger vom Ort im Raum, passt aber häufig besser zum subjektiven Empfinden des Raumklangs [71]. Die Berechnung der EDT erfolgt analog zur Nachhallzeit am genauesten über die Gleichungen (2.7) und (2.8).

Neben den bisher beschriebenen Raumparametern gibt es zahlreiche weitere, deren Relevanz aber je nach Anwendungsfall variiert. Im Falle der Musik sind das beispielsweise das Klarheitsmaß, welches die zeitliche Durchsichtigkeit beschreibt, und das Raumeindrucksmaß als Zusammenspiel von Räumlichkeit und Halligkeit [149]. Diese Parameter spielen hauptsächlich für Zuhörer in Konzerträumen und nicht bei Aufnahmen eine Rolle, weshalb sie hier nicht weiter ausgeführt werden.

2.1.2 Musiksignale

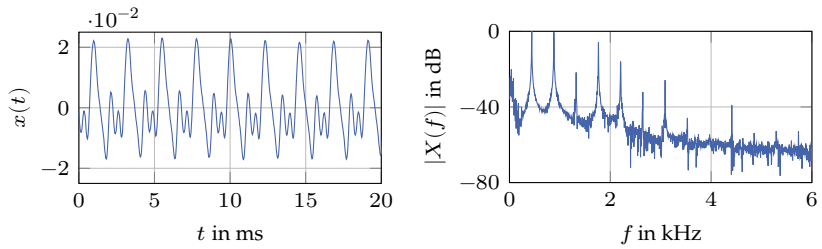
Als Unterklasse von akustischen Signalen stellen Musiksignale eindimensionale Signale dar, die zeitabhängige akustische Informationen über Musik beinhalten. Typischerweise werden diese durch Musikinstrumente und Gesangsstimmen erzeugt und mithilfe von Mikrofonen in ein elektrisches Signal umgewandelt [98].

In der westlich geprägten Musik erfolgt die Klangerzeugung meist in definierten Tönen mit fester Grundfrequenz. Die Spanne zwischen zwei Tönen mit verdoppelter Grundfrequenz wird als Oktave bezeichnet und entsprechend der gleichmäßig temperierten Stimmung in zwölf Halbtonschritte unterteilt. Zwei aufeinanderfolgende Töne besitzen dadurch immer das gleiche Frequenzverhältnis von $\sqrt[12]{2}$:1, was einem Halbtonschritt entspricht. Zur Beschreibung von Abweichungen wird ein Halbtonschritt weiter in 100 cent unterteilt. Nach DIN 1317 dient der Kammerton a' (amerikanische Schreibweise: A4) mit 440 Hz als Bezugsfrequenz der Tonskala, wobei er in vielen Orchestern etwas höher, häufig auf 442 Hz, gestimmt wird [149]. Aufgrund von veränderter Raumtemperatur und Luftfeuchte können sich die Frequenzen verschieben, sodass ein Nachstimmen notwendig ist. Durch die Verwendung von Computern hat sich neben der traditionellen Benennung der Töne das *Musical Instrument Digital Interface* (MIDI) etabliert, das den Tönen C_3 bis g^6 die Zahlen 0 bis 127 zuordnet. Damit lässt sich jede Frequenz eines MIDI-Tons i über

$$f_i = 2^{\frac{i-69}{12}} \cdot 440\,\mathrm{Hz} \tag{2.9}$$

berechnen [98]. Der Kammerton a' hat die MIDI-Nummer 69, weshalb sich die Formel auf diese Nummer bezieht.

Beispielhaft ist in Abbildung 2.2(a) der zeitliche Signalverlauf des von einer Flöte gespielten a' abgebildet, welcher den ausgeprägten periodischen Charakter illustriert. Neben der Grundfrequenz von ca. 440 Hz enthält das Zeitsignal weitere relevante Signalanteile bei ganzzahligen Vielfachen der Grundfrequenz, den sogenannten Oberschwingungen oder Harmonischen. Diese lassen sich mithilfe der Fourier-Transformation $X(f)$ des Signals $x(t)$ näher analysieren. Das daraus resultierende, auf sein Maximum normierte Amplitudenspektrum des Flötensignals ist in Abbildung 2.2(b) im logarithmischen dB-Maßstab dargestellt. Darin treten die Anteile der Grund- und Oberschwingungen deutlich als



(a) Zeitsignal. **(b)** Normiertes Betragsspektrum.

Abbildung 2.2 Zeitsignal und zugehöriges, auf sein Maximum normiertes Amplitudenspektrum des Kammertons a' einer Flötenaufnahme aus dem URMP-Datensatz [75].

schmale lokale Maxima hervor. Im Allgemeinen enthält die Grundfrequenz die größte Energie und die Intensitäten der Obertöne nehmen mit steigender Frequenz ab, was mit einzelnen Ausnahmen auch im Beispiel der Flöte zu erkennen ist. Die genaue Intensitätsverteilung auf die Grund- und Oberschwingungen charakterisiert den spezifischen Klang eines Instruments, der durch unterschiedliche Spieltechniken variieren kann [149]. Darüber hinaus tragen teilweise Inharmonizitäten in den Oberschwingungen zu dem spezifischen Instrumentenklang bei, welche einzelne Oberfrequenzen $l \in \mathbb{N}\backslash\{1\}$ der Grundfrequenz f_0 auf

$$f_l = l \cdot f_0 \sqrt{1 + B \cdot l} \tag{2.10}$$

verschieben, wobei $B \in [0, 1]$ den Inharmonizitätsfaktor darstellt [33].

Weitere wichtige Parameter zur Beschreibung von Musiksignalen sind die Lautstärke und die Noteneinhüllende. Signaltechnisch ist die Lautstärke durch die Signalamplitude bestimmt und wird häufig in Dezibel angegeben. In der menschlichen Wahrnehmung entspricht der Signalpegel aber nicht immer der empfundenen Lautstärke, weshalb dort sehr häufig die Einheit Phon verwendet wird. Sie definiert die Intensitäten im hörbaren Frequenzbereich, die als gleich laut empfunden werden. Diese sind frequenzabhängig. Die genauen Verläufe der Phon-Skala, deren höchste Empfindlichkeit bei ca. 2 kHz bis 5 kHz liegt, wurden von Fletcher und Munson [34] experimentell untersucht. Der Verlauf einzelner Töne wird häufig in die vier aufeinanderfolgenden Phasen Anschlag, Abklingen, Halten und Ausklingen unterteilt, die zusammen

die Noteneinhüllende bilden. Diese relativ einfache Beschreibung ist nach den englischen Bezeichnungen *attack*, *decay*, *sustain* und *release* als ADSR-Modell [20] bekannt und hat sich in vielen MIR-Anwendungen zur Notenerkennung und -beschreibung bewährt.

Um Melodien nachspielen zu können, werden die Töne definierter Frequenz zusammen mit der gewünschten Länge als Noten aufgeschrieben. Zusätzlich werden Instrumentierung, Rhythmik, Dynamik und Spieltechnik notiert. Eine Einführung in die Musiknotation gibt beispielsweise Vinci [146]. In dieser Arbeit geht es hauptsächlich um die Analyse von digital aufgenommenen Musiksignalen, deren Notation als unbekannt angenommen wird und somit keine Rolle spielt. Darüber hinaus hängt das aufgenommene Musiksignal nicht nur von der Notation, sondern auch von den Umsetzungen der Künstler ab.

Um eine digitale Aufnahme zu ermöglichen, wird das Musiksignal $x(t)$ äquidistant abgetastet, sodass eine zeitdiskrete Darstellung $x[n]$ des Musiksignals vorliegt. Typischerweise erfolgt diese Abtastung mit 44,1 kHz für CD-Qualität oder im Bereich zwischen 48 kHz und 96 kHz für professionelle Studioaufnahmen [98]. Unter Berücksichtigung des Abtasttheorems sind damit Frequenzen bis 22,05 kHz bzw. 24 kHz bis 48 kHz darstellbar, sodass stets der Hörbereich des Menschen enthalten ist. Die Diskretisierung der Amplitudenwerte erfolgt in der Regel mit mindestens 16 bit, wodurch 65 536 Werte möglich sind. Daraus folgt ein Signal-Rausch-Verhältnis (SNR) von ungefähr 88 dB für typische Musiksignale mit gauß- oder laplaceverteilten Amplitudenhäufigkeiten [149]. Eine feinere Quantisierung erhöht das erzielbare SNR. Im Folgenden wird ausschließlich mit digitalen Musiksignalen gearbeitet, weshalb in den weiteren Ausführungen nur noch diskrete Signale betrachtet werden.

2.2 Diskrete Zeit-Frequenz-Darstellungen

Wie in Abschnitt 2.1.2 ausgeführt, setzen sich Musiksignale meist aus Noten mit definiertem Spektrum und vorgegebener Rhythmik zusammen. Daher sind sowohl Frequenzinformationen als auch Informationen über den zeitlichen Verlauf essenziell. Eine reine Fourier-Transformation stellt dagegen nur die spektrale Information über die gesamte Beobachtungsdauer dar. Deshalb werden in diesem Kapitel zwei der für

Musiksignale wichtigsten Zeit-Frequenz-Transformationen, die Kurzzeit-Fourier-Transformation und die Constant-Q-Transformation, vorgestellt. Diese liefern Signalspektren in Abhängigkeit der Zeit und eignen sich somit als Vorverarbeitungsschritt für viele Musiksignalanalysen. Da ausschließlich diskrete Musiksignale analysiert werden, sind alle Erläuterungen auf den diskreten Fall bezogen.

2.2.1 Kurzzeit-Fourier-Transformation

Um eine zeitabhängige Analyse der Signalfrequenzen zu ermöglichen, wird das betrachtete diskrete Signal $x[n]$ zunächst mit einem Zeitfenster $w[n]$ multipliziert, dessen Signalenergie auf einen kurzen Zeitabschnitt konzentriert ist. Die Fourier-Transformation des resultierenden Signals enthält dann ausschließlich die zu diesem Zeitabschnitt vorhandenen Frequenzanteile. Durch Verschiebung des Zeitfensters der Breite N_w ist eine Analyse der in den jeweiligen Zeitintervallen vorliegenden Spektren möglich. Dieses Verfahren wird als Kurzzeit-Fourier-Transformation (englisch *short-time Fourier transform*, STFT) bezeichnet und lässt sich im Diskreten mithilfe von

$$X_{\mathrm{STFT}}[m,k] = \mathcal{F}\left\{x[n]\, w^*[n-mh]\right\} \tag{2.11}$$

$$= \sum_{n=0}^{N-1} x[n]\, w^*[n-mh]\, \mathrm{e}^{-\mathrm{j}2\pi\frac{kn}{K}} \tag{2.12}$$

berechnen [105]. Dabei stellen m und k die Indizes für die insgesamt M Zeit- und K Frequenzbins dar. Die zeitliche Auflösung der STFT erfolgt mit einer Schrittweite von h Zeitsignalwerten, welche der Fensterverschiebung entspricht und häufig als *Hop Size* angegeben wird. Daraus ergeben sich die den Indizes zugeordneten, mittleren Zeit- und Frequenzwerte

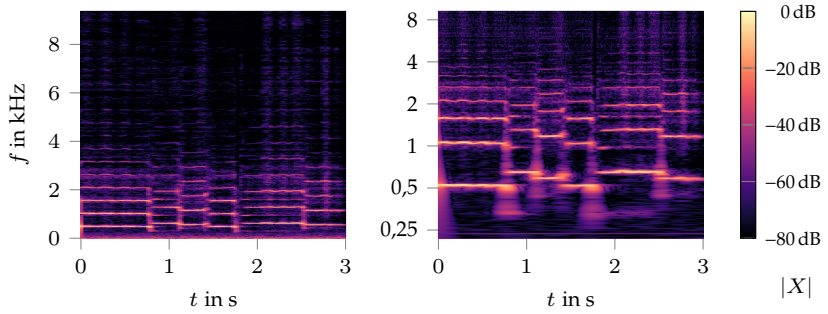
$$t_{\mathrm{STFT}}[m] = \frac{1}{f_{\mathrm{A}}}\left(mh + \frac{N_w}{2}\right) \quad \text{und} \quad f_{\mathrm{STFT}}[k] = \frac{k f_{\mathrm{A}}}{K}\,. \tag{2.13}$$

Als Analysefenster kommen zahlreiche Fensterfunktionen in Frage, die unter anderem von Harris [43] ausgiebig untersucht und miteinander verglichen wurden. In der Musiksignalverarbeitung haben sich hauptsächlich das Hamming- und das Hann-Fenster etabliert, da sie auch bei

teilweiser zeitlicher Überlappung eine einfache Signalrekonstruktion ermöglichen [39, 89]. Zur Anwendung in Gleichung (2.12) müssen die Amplitudenwerte der jeweiligen Fensterfunktion mit Nullen außerhalb der eigentlichen Fensterbreite N_w fortgeführt werden, um die geforderte Signallänge N des betrachteten Signals $x[n]$ abdecken zu können. Dies wird als *Zero-Padding* bezeichnet.

Innerhalb der mithilfe des Fensters betrachteten Zeitdauer gilt die Voraussetzung, dass der Frequenzgehalt des analysierten Signals $x[n]$ nahezu konstant bleibt [63]. Andernfalls enthält das Spektrum einen Mittelwert der dynamischen Frequenzanteile, was gerade im Falle der Musiksignalspektren mit einzelnen schmalen Maxima zu deutlichen Verschmierungen und damit zu starken Signalverfälschungen führen kann. Durch die Anwendung ausreichend kurzer Zeitfenster mit einer Dauer von ungefähr 50 ms kann diese geforderte Quasistationarität in der Musik meist eingehalten werden, da das Spektrum einer ausgehaltenen Note innerhalb dieser Dauer in der Regel annähernd konstant ist.

Die diskrete Unterteilung in M Zeit- und K Frequenzbins erfolgt in beiden Dimensionen äquidistant, da sowohl die zeitliche Schrittweite h als auch die Breite des Analysefensters N_w, welche die maximal zu erreichende spektrale Auflösung beschränkt, konstant sind. Folglich sind die Grundfrequenz und die zugehörigen Harmonischen in der Betragsdarstellung der Kurzzeit-Fourier-Transformation durch Linien in gleichem Frequenzabstand zu erkennen. Dies wird in Abbildung 2.3(a) anhand der Betragsdarstellung des Beispielsignals einer Flötenaufnahme aus dem URMP-Datensatz [75] veranschaulicht. Da die Linien das Spektrum eines ausgehaltenen Tons repräsentieren, wird die Annahme der Quasistationarität für das hier verwendete Hann-Fenster mit $N_w = 4096$ und 48 kHz Abtastrate bestätigt. Auch für leichte Variationen der Frequenz, wie z. B. zu Beginn bis etwa 0,7 s, werden klare Frequenzverläufe angezeigt, nur an den Tonübergängen zeigt sich ein Verschmieren der Spektren. Darüber hinaus werden die schon in Abbildung 2.2(b) festgestellten Charakteristika einer prägnanten Grundfrequenz und der starken Intensitätsabnahme für höhere Oberschwingungen verdeutlicht. Die Signalenergie der Flötenaufnahme ist hauptsächlich im Bereich tiefer Frequenzen konzentriert, weshalb andere Zeit-Frequenz-Transformatio-



(a) Betrag der Kurzzeit-Fourier-Transformation.

(b) Betrag der Constant-Q-Transformation.

Abbildung 2.3 Verschiedene Zeit-Frequenz-Darstellungen des gleichen Ausschnitts einer Flötenaufnahme (18_Nocturne) aus dem URMP-Datensatz [75].

nen wie die in Abschnitt 2.2.2 beschriebene Constant-Q-Transformation den tiefen Frequenzbereich genauer auflösen.

Über die Wahl der STFT-Parameter kann die Auflösung ebenfalls gezielt verändert werden. So folgt aus einer größeren Fensterbreite eine feinere Frequenzauflösung, da durch die Fourier-Transformation mehr Frequenzbins für das identische Frequenzintervall berechnet werden können. Dies geht allerdings auf Kosten der Zeitauflösung, weil das breitere Fenster mehr diskrete Signalwerte umfasst. Wegen dieser sogenannten Unschärferelation muss folglich eine für die jeweilige Anwendung geeignete Balance zwischen Zeit- und Frequenzauflösung gefunden werden.

Aus dem komplexwertigen Ergebnis $X_{\mathrm{STFT}}[m,k]$ der STFT lässt sich das zugehörige diskrete Zeitsignal wieder vollständig mithilfe eines zum Analysefenster $w[n]$ passenden Synthesefensters $\tilde{w}[n]$ rekonstruieren. Das rücktransformierte Zeitsignal

$$\hat{x}[n] = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} X_{\mathrm{STFT}}[m,k]\, \tilde{w}[n-mh]\, \mathrm{e}^{\mathrm{j}2\pi \frac{kn}{K}} \tag{2.14}$$

entspricht allerdings nur dann exakt dem ursprünglichen Zeitsignal $x[n]$, wenn Analyse- und Synthesefenster zueinander biorthogonal sind. Die Herleitung dieser Bedingung sowie eine Anleitung zur Berechnung der

passenden Synthesefunktion aus einem vorgegebenen Analysefenster wird z. B. von Puente León und Jäkel [105] beschrieben.

Auf Basis der STFT-Transformierten haben sich mehrere Darstellungen für Musiksignale entwickelt, die bestimmte Signaleigenschaften durch Weiterverarbeitung stärker hervorheben. Beispiele hierfür sind das Spektrogramm mit logarithmischer Frequenzachse, das die Frequenzbins der ursprünglichen STFT neu anhand der logarithmischen Skala zusammenfasst, sowie das Chromagramm, welches vorkommende Musiknoten unabhängig von ihrer Oktave repräsentiert [98]. Eine weitere, häufig verwendete Darstellung ist das Mel-Spektrogramm, das die Frequenzbins der STFT in die psychoakustische Mel-Skala [130] überführt, die auf der für tiefe Frequenzen annähernd linearen und für hohe Frequenzen logarithmischen Frequenzwahrnehmung des menschlichen Gehörs basiert. Dadurch lassen sich in manchen Teilbereichen der MIR vergleichbare Ergebnisse wie mit der STFT erzielen, obwohl die Anzahl an Mel-Frequenzbins im Vergleich zur STFT drastisch reduziert ist [11]. Die Auflösung der STFT kann durch diese Verfahren allerdings nicht verbessert werden. Des Weiteren haben alle diese Darstellungsformen das Problem, dass eine Rücktransformation in das zugehörige Zeitsignal durch die Neuzuordnung der Frequenzbins nicht möglich ist.

2.2.2 Constant-Q-Transformation

Eine in der Verarbeitung und Darstellung von Musiksignalen weit verbreitete Zeit-Frequenz-Analyse ist die Constant-Q-Transformation (CQT). Sie stellt einen Spezialfall der Wavelet-Transformation dar, deren frequenzabhängige Zeit- und Frequenzauflösung sehr gut zur exponentiellen Einteilung der Tonfrequenzen aus Gleichung (2.10) passt, weil die Verdopplung der Tonfrequenz pro Oktave eine deutlich genauere Frequenzauflösung für tiefe Frequenzen erfordert. Die CQT besitzt eine logarithmische Frequenzskala, sodass das Verhältnis

$$Q = \frac{f[k]}{\Delta f[k]} = \frac{f[k] \cdot N_w[k]}{f_\mathrm{A}} \tag{2.15}$$

aus mittlerer Frequenz $f[k]$ und zugehöriger Bandbreite $\Delta f[k]$ konstant ist. Diese namensgebende Eigenschaft der CQT wird nach Gleichung (2.15) erreicht, indem die Breite des Analysefensters N_w invers

zu der zu analysierenden Frequenz mit Index k variiert [6]. Durch die logarithmische Frequenzskala und das konstante Q sind die Abstände zwischen den Grundfrequenzen des gleichen Tonintervalls in der CQT unabhängig von der absoluten Lage der Grundfrequenz immer gleich groß. Darüber hinaus bilden die Grund- und Oberschwingungen im Frequenzbereich ein festes, frequenzunabhängiges Muster, wodurch Musiktöne und Instrumentenspektren gut zu identifizieren sind.

Analog zur allgemeinen Wavelet-Transformation wird die frequenzabhängige Variation des Analysefensters mithilfe einer Skalierung umgesetzt. Im Diskreten lässt sich das skalierte Fenster durch

$$w_k[n] = \frac{1}{N_w[k]} \cdot w\left(\frac{n}{N_w[k]}\right) \tag{2.16}$$

beschreiben, wobei die ursprüngliche Fensterfunktion $w(t)$ in dieser Gleichung kontinuierlich vorliegt und sich über den Bereich $t \in [0, 1]$ erstreckt [117]. Um mit den skalierten Fenstern aller Frequenzindizes k eine Analyse definierter (mittlerer) Zeitpunkte vornehmen zu können, müssen die Fenster $w_k[n]$ die gleiche mittlere Zeit besitzen. Deshalb werden die skalierten Fenster zeitlich so verschoben, dass sie diese Anforderung erfüllen. In der Regel werden sie um Null zentriert. Anschließend werden sie, wie schon bei der STFT, mithilfe von *Zero-Padding* auf die zur Analyse des gesamten Signals $x[n]$ erforderliche Länge erweitert.

Die Zeitabhängigkeit der Frequenzspektren wird analog zur STFT durch zeitliche Verschiebung der skalierten Analysefenster mit konstanter *Hop Size* berücksichtigt. Im Unterschied zur STFT fließt diese Zeitverschiebung in der CQT aufgrund der Skalierung zusätzlich in die Exponentialfunktion mit ein, woraus sich die Formel

$$X_{\mathrm{CQT}}[m, k] = \sum_{n=0}^{N-1} x[n]\, w_k^*[n - mh]\, \mathrm{e}^{-\mathrm{j}2\pi \frac{Q \cdot (n - mh)}{N_w[k]}} \tag{2.17}$$

zur Berechnung der diskreten CQT für die gesamte Signallänge N ergibt [118]. Dabei repräsentieren m und k auch hier die Indizes für die insgesamt M Zeit- und K Frequenzbins, deren mittlere Zeit- und Frequenzwerte durch

$$t_{\mathrm{CQT}}[m] = \frac{1}{f_{\mathrm{A}}} \left(mh + \Delta N_{w,0}\right) \quad \text{und} \quad f_{\mathrm{CQT}}[k] = f_{\min} \cdot 2^{\frac{k}{b}} \tag{2.18}$$

gegeben sind. Der konstante Term $\Delta N_{w,0}$ beschreibt die initiale Zeitverschiebung der skalierten Fensterfunktionen und ist folglich Null, falls die Fenster $w_k[n]$ um Null zentriert sind. Die mittlere Frequenz ist abhängig von der Anzahl an Frequenzbins pro Oktave b und der kleinsten in der CQT berücksichtigten Frequenz $f_{\min}$, welche gleichbedeutend mit dem Frequenzwert für $k = 0$ ist. Sowohl b als auch $f_{\min}$ sind wählbare Designparameter der CQT. Durch die Wahl der Frequenzbins pro Oktave über den Parameter b wird das konstante

$$Q = \frac{1}{2^{\frac{1}{b}} - 1} \tag{2.19}$$

vorgegeben [117], was mithilfe der Gleichung (2.15) die für ein konstantes Q notwendigen Fensterbreiten $N_w[k]$ in Abhängigkeit der mittleren Frequenzen $f[k]$ festlegt.

In der ursprünglich von Brown [6] vorgeschlagenen Anwendung der CQT auf Musiksignale wurde als skalierte Fensterfunktion ein modifiziertes Hamming-Fenster verwendet. Die Berechnung der Constant-Q-Transformierten kann dabei effizient als Multiplikation in der Frequenzdomäne ausgeführt werden, indem die skalierten Fenster vorab als Kerne der CQT definiert und mithilfe der schnellen Fourier-Transformation (FFT) in den Frequenzbereich transformiert werden [7]. Die transformierten Kerne bilden dann eine Art Filterbank für die CQT. Ein großer Nachteil dieser Implementierung ist die fehlende Invertierbarkeit, wodurch die ursprünglichen Signale nicht aus der CQT rekonstruierbar sind. Dieser Nachteil wird durch die Anwendung einer alternativen Filterbank aus nichtstationären Gabor-Frames behoben, die zur fehlerfreien Rücktransformation duale Frames nutzt [49]. Der zugehörige Algorithmus wird auch als nichtstationäre Gabor-Transformation (NSGT) bezeichnet und ermöglicht mithilfe der FFT ebenfalls eine effiziente Berechnung über die Frequenzdomäne. Durch eine abschnittsweise CQT-Berechnung, die aufgrund der zeitlichen Abgeschlossenheit der ursprünglichen Analysefenster und dem *Zero-Padding* möglich ist [118], kann damit sogar eine Verarbeitung in Echtzeit umgesetzt werden.

Für das Beispielsignal einer Flötenaufnahme ist die Betragsdarstellung der CQT mit NSGT-Implementierung, $b = 48$ und 48 kHz Abtastrate in Abbildung 2.3(b) dargestellt. Ausgehaltene Töne sind als Linien hoher Energie an den jeweiligen Grund- und Oberfrequenzen zu erkennen.

Diese Linien sind im immer gleichen Muster mit abnehmendem Abstand zwischen den Harmonischen angeordnet, was an der logarithmischen Frequenzskala liegt. Verglichen mit der in Abbildung 2.3(a) gezeigten Darstellung der STFT wird die für tiefe Frequenzen um ein Vielfaches genauere Frequenzauflösung deutlich, da z. B. der Frequenzbereich bis 2 kHz im Falle der CQT mehr als die Hälfte der Frequenzachse, für die STFT aber weniger als ein Viertel einnimmt. Die Signalenergie verteilt sich in der CQT-Darstellung besser über den gesamten Diagrammbereich, weshalb die CQT für die Analyse von Musikspektren sehr gut geeignet ist. Im Gegensatz dazu ist die klassische Wavelet-Transformation für so hohe Frequenzauflösungen nicht geeignet [6, 118]. Die Zeitauflösung der CQT ist linear und somit vergleichbar zur STFT.

Wie alle hier betrachteten Zeit-Frequenz-Darstellungen unterliegt auch die CQT der Zeit-Frequenz-Unschärfe. Das wird unter anderem an den Tonwechseln in Abbildung 2.3(b) deutlich, die einen leicht perkussiven Charakter haben und deshalb Anteile fast im gesamten Frequenzbereich besitzen. Aufgrund der Unschärferelation weiten sich diese eigentlich in einem sehr kurzen Zeitintervall lokalisierten Anteile für kleine Frequenzen zeitlich aus, sodass in der Betragsdarstellung nach unten breiter werdende, annähernd dreieckige Strukturen um die jeweiligen Zeitpunkte herum entstehen. Diese kommen durch die breiten Analysefenster zustande, welche für die Betrachtung tiefer Frequenzen notwendig sind. Dadurch decken die Fenster einen großen Zeitbereich ab, um einen sehr schmalen Frequenzbereich zu analysieren.

Um die zeitliche Unschärfe für sehr tiefe Frequenzen einzuschränken, kann das sonst als konstant geforderte Q für Frequenzen unterhalb einer definierten Schwelle mithilfe eines zusätzlichen Faktors reduziert werden. Die höhere zeitliche Auflösung der betroffenen tiefen Frequenzen hat dann allerdings auch eine reduzierte Frequenzauflösung in diesem Bereich zur Folge. Angelehnt an die menschliche Wahrnehmung kann die Schwelle bei 500 Hz gewählt werden, da für die Frequenzauflösung des menschlichen Gehörs erst ab ungefähr dieser Frequenz ein konstantes Q angenommen werden kann [118]. Eine Alternative zur CQT für die menschliche Wahrnehmung ist die ERBlet-Transformation [100], deren Werte für Q nicht konstant sind, sondern der spektralen Auflösung des menschlichen Gehörs entsprechen.

2.3 Künstliche neuronale Netze

Schon seit der Erfindung programmierbarer Computer faszinierte Menschen die Idee von intelligenten Maschinen und Computern, die eigenständig Aufgaben ausführen, Entscheidungen treffen und lernen [37]. Das dadurch beschriebene Feld der künstlichen Intelligenz (KI) spielt durch die zunehmende Automatisierung eine immer größere Rolle. Viele KI-Algorithmen basieren auf menschlicher Expertie, die in definierte mathematische Regeln überführt und somit fest programmiert wird. Im KI-Teilbereich des maschinellen Lernens (ML) werden dagegen datengetriebene Ansätze verfolgt, deren Funktionen durch die im Laufe eines Trainings generierten Erfahrungen automatisch verbessert werden [95]. Für dieses schrittweise Lernen sind in der Regel große Datenmengen erforderlich, um allgemeingültige und nicht nur sehr spezifische Zusammenhänge des betrachteten Themengebiets zu erlernen. Die komplexen Informationsgehalte großer Datensätze können oft nur mithilfe tiefer, umfangreicher Strukturen abgebildet werden, weshalb sich das sogenannte *Deep Learning* (DL) etabliert hat [19]. Dabei werden tiefe Modelle eingesetzt, die meist als künstliche neuronale Netze (KNN) mit vielen Schichten aufgebaut sind. Als Teilbereich des maschinellen Lernens stellen KNN Netzwerke mit lernbaren Parametern und Gewichten dar, welche die neuronalen Netze mit komplex verschalteten Nervenzellen in den Gehirnen von Menschen und Tieren nachbilden [29]. Durch die zunehmende Verfügbarkeit leistungsstarker Ressourcen haben sich tiefe KNN in den letzten Jahren in vielen Anwendungsbereichen durchgesetzt.

Abhängig von der Art der geforderten Ausgangsdaten werden KNN in Klassifikations- oder Regressionsnetze eingeteilt. Klassifikationsnetze erfüllen die qualitative Aufgabe der Klassifikation, in der die Ausgangswerte einer begrenzten Menge diskreter Klassen zugeordnet werden. Im Gegensatz dazu können die Ausgänge eines Regressionsnetzes beliebige numerische Werte innerhalb vorgegebener Grenzen abdecken, weshalb die Regression eine quantitative Aufgabe darstellt.

Eine weitere Kategorie zur Einteilung der KNN ist ihre Grundarchitektur. Die wichtigsten KNN-Bausteine und die daraus aufgebauten Architekturen werden in den Abschnitten 2.3.1 und 2.3.2 vorgestellt. Anschließend wird kurz auf die Grundprinzipien des Trainings eingegangen. Detailliertere Einführungen in die Gebiete des DL sowie der

künstlichen neuronalen Netze sind in zahlreichen Büchern wie z. B. von Calin [8] oder Goodfellow et al. [37] zu finden. Speziell für Musiksignale haben Choi et al. [12] einen Überblick über die wichtigsten Netze und Herangehensweisen zusammengestellt.

2.3.1 Bausteine

Künstliche neuronale Netze basieren wie die neuronalen Netze biologischer Gehirne auf der Verknüpfung einer Vielzahl von kleinen Schaltelementen, den Neuronen, und wurden schon im Jahre 1943 von McCulloch und Pitts [88] eingeführt. Die Neuronen besitzen einen Ausgang y sowie I Eingänge x_i und werden mathematisch durch

$$y = \sigma\left(\sum_{i \in I} \alpha_i x_i + \beta\right) \tag{2.20}$$

beschrieben, wobei α_i die trainierbaren Gewichte und β den *Bias* des Neurons darstellen. Um nichtlineare und damit auch komplexe Beziehungen abbilden zu können, muss die Aktivierungsfunktion σ eine nichtlineare Funktion sein. Inspiriert durch die biologischen Neuronen wurden, vor allem in den Anfängen der KNN, die logistische Sigmoid-Funktion und der Tangens hyperbolicus

$$\sigma_{\text{Sigmoid}}(x) = \frac{1}{1 + \mathrm{e}^{-x}} \qquad \text{bzw.} \qquad \sigma_{\tanh}(x) = \frac{\mathrm{e}^{x} - \mathrm{e}^{-x}}{\mathrm{e}^{x} + \mathrm{e}^{-x}} \tag{2.21}$$

als Aktivierungsfunktionen eingesetzt [25]. Mittlerweile wurden sie durch die *Rectified Linear Unit* (ReLU) abgelöst, deren Aktivierung

$$\sigma_{\text{ReLU}}(x) = \max(0, x) \tag{2.22}$$

negative Eingangswerte zu Null setzt. Gegenüber den ursprünglichen Funktionen besitzt die ReLU-Funktion den Vorteil, dass sie im positiven Wertebereich unbeschränkt ist und damit nie in die Sättigung geht. Darüber hinaus besitzt ReLU eine deutlich kürzere Berechnungsdauer [29, 36]. Neben den drei vorgestellten, elementaren Aktivierungsfunktionen gibt es viele weitere, die z. B. Dubey et al. [25] zusammengestellt haben.

Um Elemente mit mehr als einem Ausgang zu bilden, werden mehrere parallele Neuronen in einer sogenannten Schicht gruppiert. Die

Neuronen einer Schicht sind untereinander nicht verknüpft. Ein KNN entsteht aus der Reihenschaltung mehrerer Schichten, sodass die Neuroneneingänge einer Schicht von den Ausgängen der vorherigen Schicht gespeist werden. Mathematisch kann das Verhalten jeder Schicht als Matrixmultiplikation

$$\mathbf{y} = \sigma\left(\mathbf{A}\mathbf{x} + \boldsymbol{\beta}\right) \tag{2.23}$$

beschrieben werden, wobei die Matrix $\mathbf{A}$ alle Gewichte und der Vektor $\boldsymbol{\beta}$ die *Bias*-Werte der einzelnen Neuronen der entsprechenden Schicht enthält. Falls jedes Neuron einer Schicht alle Ausgänge der vorausgehenden verarbeitet, wird diese Schicht als voll verbunden bezeichnet.

Für große KNN sind voll verbundene Schichten oft ungeeignet, da sie durch die Verbindungen zwischen allen Neuronen aufeinanderfolgender Schichten sehr viele freie Parameter besitzen. Deshalb haben sich Faltungsschichten etabliert, in denen die Matrixmultiplikation der voll verbundenen Schicht durch eine Faltung ersetzt wird [37]. Dabei werden die Eingänge mit einem Filterkern gefaltet, der immer nur einen Teil der Eingänge berücksichtigt und mit definierter Schrittweite über die räumliche Dimension des gesamten Eingangsbereichs verschoben wird. Die Faltungsschicht besteht folglich nicht mehr aus unabhängigen Neuronen, sondern die lernbaren Gewichte des Filterkerns stellen die gemeinsamen, freien Parameter der Schicht dar. Typischerweise enthält eine Faltungsschicht nicht nur einen, sondern mehrere Faltungskerne, um gleichzeitig mehrere Merkmale zu extrahieren, sodass die Formel

$$\mathbf{Y}^{(V)} = \sigma\left(\sum_{u \in U} \mathbf{C}^{(Vu)} * \mathbf{X}^{(u)} + \mathbf{B}^{(V)}\right) \tag{2.24}$$

für mehrdimensionale Ein- und Ausgangstensoren $\mathbf{X}$ bzw. $\mathbf{Y}$ und die Faltungskerne $\mathbf{C}$ gilt, wobei U und V die Anzahl der verschiedenen Ein- und Ausgangsmerkmale repräsentieren. Obwohl vergleichbare Ergebnisse wie mit voll verbundenen Schichten erzielt werden, fällt die Parameteranzahl der Faltungsschichten deutlich geringer aus, wodurch sie wesentlich schneller zu trainieren sind [8]. Gerade für zweidimensionale Daten wie Bilder sind Faltungsschichten sehr gut geeignet, da sie nahe örtliche Abhängigkeiten berücksichtigen [1]. Am Ausgang der Schicht liegen in diesem Fall ebenfalls zweidimensionale Merkmale vor, die häufig als Merkmalskarten bezeichnet werden.

In Kombination mit Faltungsschichten werden meistens auch sogenannte *Pooling*-Schichten eingesetzt. Sie reduzieren die Dimensionen der eingehenden Merkmalskarten, indem sie kleine rechteckförmige Ausschnitte der Karte in einem Wert zusammenfassen. Als Reduktionsmethode haben sich die Mittelwertbildung (*Average-Pooling*), aber vor allem die Extraktion des Maximums (*Max-Pooling*) als vorteilhaft erwiesen. Durch die Verwendung von *Pooling* ist das KNN invariant gegenüber kleinen Verschiebungen und Störungen [73]. Des Weiteren reduziert es kleine Details und abstrahiert die eingehenden Merkmale, wodurch nachfolgende Schichten größere Zusammenhänge berücksichtigen können.

Sequenzielle Daten besitzen eine hohe Abhängigkeit innerhalb der einzelnen Datenpunkte, weshalb für sie eine schrittweise Verarbeitung in der vorgegebenen Reihenfolge geeignet ist. Eine solche sequenzielle Struktur bieten rekurrente Schichten, deren Neuronen neben den oben beschriebenen Verknüpfungen zu Neuronen anderer Schichten jeweils eine rekursive Verbindung enthalten, über die ihr Ausgang im nächsten Schritt wieder als Eingang desselben Neurons anliegt. Kern der rekurrenten Schicht ist ihr Zustand

$$\mathbf{z}_n = \sigma\left(\mathbf{A}_{\mathbf{zz}}\,\mathbf{z}_{n-1} + \mathbf{A}_{\mathbf{zx}}\,\mathbf{x}_n + \boldsymbol{\beta}_{\mathbf{x}}\right) \tag{2.25}$$

zum Schritt n, der über die Einbeziehung des Vorgängerzustands $\mathbf{z}_{n-1}$ Informationen der bisherigen Sequenz enthält und als „Gedächtnis“ der rekurrenten Schicht interpretiert werden kann. Der Ausgang

$$\mathbf{y}_n = \mathbf{A}_{\mathbf{yz}}\,\mathbf{z}_n + \boldsymbol{\beta}_{\mathbf{y}} \tag{2.26}$$

hängt direkt vom aktuellen Zustand der Schicht ab [8]. Durch die Verwendung gemeinsamer Gewichte für alle Daten der Sequenz haben rekurrente Schichten eine variable Eingangslänge [50].

Das Hauptproblem rekurrenter Schichten ist die direkte Rückkopplung des Zustands, weil Abhängigkeiten über eine große Zahl von Schritten nur schwer behalten werden können. Darüber hinaus kann im Training der Optimierungsgradient durch die lange sequenzielle Abhängigkeit sehr klein werden, was als *Vanishing Gradient Problem* bekannt ist [73]. Diese Probleme werden durch die von Hochreiter und Schmidhuber [48] vorgestellten *Long Short-Term Memory*- (LSTM-) Zellen gelöst, welche langfristige Abhängigkeiten durch steuerbare Speicherzellen lernen und

abbilden. Der Inhalt jeder Speicherzelle kann mithilfe eines logischen Schaltelements, dem sogenannten *Forget Gate*, teilweise gelöscht oder mithilfe eines zweiten Schaltelements, dem *Update Gate*, verändert werden. Als drittes Schaltelement ist das *Output Gate* für den zugehörigen LSTM-Zustand $\mathbf{z}_n$ verantwortlich. Alle Werte der drei Schaltelemente werden mithilfe von Gleichung (2.25) berechnet, wobei $\mathbf{z}_n$ durch den entsprechenden Wert und die Gewichte durch die des jeweiligen Schaltelements ersetzt werden [8]. Eine Weiterentwicklung des LSTM ist die *Gated Recurrent Unit* (GRU), welche auch als Vereinfachung des LSTM angesehen werden kann, da die GRU ohne separate Speicherzelle auskommt [1]. Anstatt der drei LSTM-Schaltelemente besitzt die GRU mit dem *Update Gate* und dem *Reset Gate* nur zwei Schaltelemente. Im direkten Vergleich zeigen LSTM und GRU vergleichbare Ergebnisse, wobei die GRU aufgrund der geringeren Parameteranzahl einfacher zu implementieren und etwas schneller zu trainieren ist [14]. Für sehr große Datensätze hat dagegen die LSTM-Schicht wegen ihrer komplexeren Struktur leichte Vorteile.

2.3.2 Architekturen

Aus den in Abschnitt 2.3.1 vorgestellten Elementen lassen sich diverse KNN-Strukturen aufbauen. Dazu werden die eingesetzten Schichten meist hintereinander angeordnet und die Eingänge jeder Schicht mit den Ausgängen der vorausgehenden Schicht verbunden. Folglich besitzt ein KNN in der Regel eine Eingangsschicht, an der die Eingangsdaten eingespeist werden, eine Ausgangsschicht, welche die Ausgangsdaten ausgibt, und mehrere verborgene Schichten, in denen die Daten ohne Einblick des Nutzers verarbeitet werden.

Im einfachsten Fall wird ein KNN ausschließlich aus voll verbundenen Schichten gebildet. Durch die Entwicklung von Faltungsschichten haben sich in vielen Anwendungen wie der Bildverarbeitung künstliche neuronale Faltungsnetze (englisch *convolutional neural networks*, CNN) etabliert, die mindestens eine Faltungsschicht enthalten [50]. Wie schon in Abschnitt 2.3.1 ausgeführt, werden durch die Faltungsschichten mit vergleichsweise wenigen Parametern unterschiedliche ortsunabhängige, lokale Merkmale extrahiert. Mehrere aufeinanderfolgende Faltungsschichten und nachgeschaltete *Pooling*-Schichten ermöglichen die Extraktion größerer und abstrakterer Merkmale oder Muster. Diese können

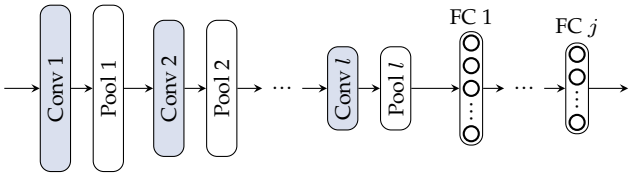


Abbildung 2.4 Schema einer klassischen CNN-Architektur mit l Faltungs- (Conv) und *Pooling*-Schichten (Pool) zur Merkmalsextraktion sowie j voll verbundenen Schichten (FC).

mithilfe von voll verbundenen Schichten am Ende des KNN in Beziehung gesetzt und global ausgewertet werden, woraus sich die klassische, in Abbildung 2.4 dargestellte CNN-Architektur ergibt. Sie ist vor allem für Klassifikationsaufgaben gut geeignet.

Grundsätzlich ist die Verwendung von kleinen Faltungskernen und vielen hintereinandergeschalteten Faltungsschichten vorteilhaft, weil dadurch kleine Details berücksichtigt und gleichzeitig in tieferen Stufen hohe Abstraktionsgrade erreicht werden. Aus diesen Überlegungen resultiert die VGG-Architektur [124] mit 11 bis 19 Faltungsschichten. Für noch tiefere Architekturen tritt das schon in Abschnitt 2.3.1 beschriebene Problem des *Vanishing Gradient* auf, das eine effiziente Optimierung eines sehr tiefen Netzes unmöglich macht. Dieses Problem kann durch die Einführung von sogenannten *Residual*-Modulen verhindert werden, die mehrere Schichten durch eine parallele Verbindung umgehen. Der Wert dieser *Skip*-Verbindung, der eine Kopie des Moduleingangs darstellt, wird zum Ausgang der letzten umgangenen Schicht addiert. Dadurch werden in den umgangenen Schichten nur die Abweichungen von den Eingangswerten des *Residual*-Moduls betrachtet und das gesamte Netz verarbeitet die Daten iterativ [1]. Zur Veranschaulichung ist in Abbildung 2.5 ein *Residual*-Modul mit drei Faltungsschichten dargestellt. Die vierte, gestrichelt dargestellte Faltungsschicht in der *Skip*-Verbindung wird nur verwendet, wenn die Dimensionen von Moduleingang und -ausgang nicht übereinstimmen. *Residual*-Module ermöglichen sehr tiefe KNN wie z. B. die ResNet-Architektur [45] mit 152 Faltungsschichten.

Im Falle sequenzieller Daten werden häufig rekurrente Schichten eingesetzt, weshalb die zugehörigen KNN als rekurrente neuronale Netze (RNN) bezeichnet werden. Ihre Architektur ist meist ähnlich zu der klas-

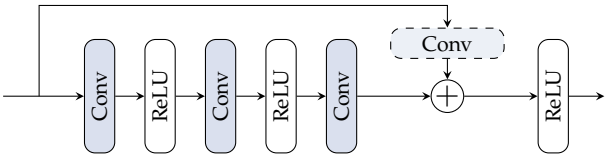


Abbildung 2.5 Struktur eines *Residual*-Moduls mit drei Faltungsschichten (Conv) und separat aufgeführter ReLU-Aktivierungsfunktionen.

sischen Architektur in Abbildung 2.4, wobei die Faltungsschichten durch rekurrente ersetzt werden. Verzweigtere Baumstrukturen reduzieren die Tiefe und steigern die Lernfähigkeit langfristiger Abhängigkeiten. Diese Architekturen werden in Abgrenzung zu RNN auch als rekursive neuronale Netze bezeichnet [37]. Sind die vollständigen Eingangssequenzen von Beginn an bekannt, verbessern bidirektionale RNN [120] die erzielten Ergebnisse, da sequenzielle Abhängigkeiten sowohl in Vorwärts- als auch in Rückwärtsrichtung gelernt werden. Seit Einführung der LSTM-Zellen werden rekurrente Schichten in RNN-Architekturen meistens durch LSTM-Schichten ersetzt, weil die dadurch entstehenden LSTM-Netze wirkungsvoller und konventionellen RNN überlegen sind [73]. Auch LSTM- oder dazu verwandte Zellen wie GRU können bidirektional umgesetzt werden, falls die vollständigen Sequenzen bekannt sind.

In den meisten Fällen sind die sequenziellen Daten Zeitsignale. Liegen am Eingang zeitabhängige Signale an und werden am Ausgang ebenfalls zeitabhängige Signale geschätzt, sind die zugehörigen KNN sogenannte *Sequence-to-Sequence*-Modelle [134]. Sie können als RNN und ebenfalls als CNN umgesetzt sein. Eine weitere Architektur zur Verarbeitung zeitabhängiger Signale bieten die Transformer [141], welche mithilfe des *Attention*-Merkmals spezielle Aufmerksamkeit auf bestimmte Eingangsdaten und gelernte Zusammenhänge legen. Sie sind sehr erfolgreich in Übersetzungsaufgaben oder der Textgenerierung [29].

Viele KNN komprimieren die anliegenden Eingangsdaten durch die strukturell vorgegebene Extraktion weniger Merkmale. Diese Eigenschaft der Datenkompression ist Hauptziel der Autoencoder-Architektur [47], die aus zwei Teilnetzen besteht. Der erste Teil ist der sogenannte Encoder, welcher die Eingangsdaten in mehreren Schichten immer weiter komprimiert. Ein Beispiel für die Umsetzung mit Faltungsschichten ist der

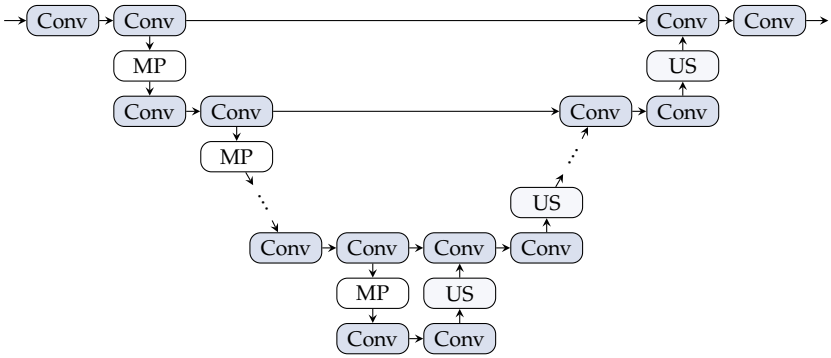


Abbildung 2.6 Schema einer UNet-Architektur aus Faltungs- (Conv), *Max-Pooling*- (MP) und *Upsampling*-Schichten (US).

vordere Teil der klassischen CNN-Architektur in Abbildung 2.4. Nach dem Encoder liegen die Daten in ihrer komprimiertesten Form, der latenten Repräsentation, vor. Anschließend werden sie im zweiten Teil, dem Decoder, wieder auf ihre ursprünglichen Dimensionen gebracht. Die Decoder-Architektur ist deshalb oft sehr ähnlich zur gespiegelten Encoder-Architektur. Klassische Autoencoder versuchen die Eingangsdaten am Ausgang exakt zu rekonstruieren. Abwandlungen wie der *Denoising* Autoencoder [145] schätzen die entrauschten Eingangsdaten.

Ähnlich wie bei Autoencodern wird die stufenweise Komprimierung und Dekomprimierung der Daten auch in der UNet-Architektur [111] eingesetzt, deren Schema in Abbildung 2.6 illustriert ist. Zur Datenkomprimierung werden in diesem Beispiel je zwei Faltungsschichten und eine *Max-Pooling*-Schicht pro Stufe verwendet. Im Decoder-Teil sind *Upsampling*-Schichten für die Dekomprimierung zuständig, welche spezielle Faltungsschichten repräsentieren, die aus einem komprimierten Wert mehrere Werte der nächsten Stufe berechnen. Die Anzahl der Faltungsschichten pro Stufe kann beliebig gewählt werden. Um das Problem des *Vanishing Gradient* zu umgehen, werden analog zu den *Residual*-Modulen *Skip*-Verbindungen zwischen Encoder- und Decoder-Schichten der gleichen Stufe eingesetzt. Dabei kann die Zusammenführung der *Upsampling*-Ergebnisse mit der *Skip*-Verbindung mithilfe einer Addition oder durch Zusammenfügen der Daten erfolgen.

2.3.3 Training

Ein Hauptbestandteil des Entwicklungsprozesses von ML-Algorithmen ist das Training. Nachdem die Modellstruktur definiert wurde, werden die veränderbaren Parameter und Gewichte des Modells während des Trainings hinsichtlich der vorgegebenen Trainingsdaten optimiert. Folglich ist die Auswahl geeigneter Trainingsdatensätze essenziell und hat einen großen Einfluss auf das gelernte Ergebnis. Gerade für tiefe Netzstrukturen wie bei KNN werden große Datenmengen benötigt, um die Modellgewichte hinsichtlich der gewünschten Aufgabe trainieren zu können. Falls die Trainingsdatensätze nicht genügend Daten umfassen, wird häufig eine Datenerweiterung (englisch *data augmentation*) durchgeführt [12]. Diese Erweiterung verändert bestehende Trainingsdaten meist so, dass sie zwar verfälscht sind und somit leicht andere Informationen enthalten, ihre Haupteigenschaften im Sinne der gewünschten Aufgabe aber erhalten bleiben, und fügt sie dann dem bisherigen Datensatz hinzu.

Ein Training erfolgt in mehreren Epochen, in der typischerweise jeder Datenpunkt des Datensatzes einmal verwendet wird. Jede Epoche ist wiederum in mehrere Teilschritte unterteilt, die jeweils eine Optimierung mit nur einer Teilmenge der Trainingsdaten (*Batch*) durchführt. Dabei minimiert die Optimierung eine vorgegebene Kostenfunktion J, wie z. B. den mittleren quadratischen Fehler (englisch *mean squared error*, MSE)

$$J_{\mathrm{MSE}} = \frac{1}{L} \sum_{i=1}^{L} \left(y_i - \hat{y}_i \right)^2 \tag{2.27}$$

oder die Kreuzentropie (englisch *cross-entropy*, CE)

$$J_{\mathrm{CE}} = - \sum_{i=1}^{L} y_i \cdot \log \left(\hat{y}_i \right) \tag{2.28}$$

der L Datenpunkte mit den Zielwerten y_i und ihrer Schätzungen $\hat{y}_i$ [55]. Der MSE wird sehr häufig als Kostenfunktion für Regressionsaufgaben eingesetzt. Zur Lösung des Optimierungsproblems haben sich im Bereich der KNN hauptsächlich der stochastische Gradientenabstieg (englisch *stochastic gradient descend*, SGD) [37] und der Adam-Optimierer [66], eine Weiterentwicklung des SGD, etabliert. Sie benötigen nicht den richtigen, sondern nur eine Schätzung des Gradienten, der mithilfe jedes *Batches*

approximiert wird. Die Optimierung der einzelnen Modellgewichte wird mithilfe des *Backpropagation*-Algorithmus erzielt, indem die Abweichungen der Zielwerte von den Schätzungen vom Ausgang aus rückwärts durch das Netz propagiert werden. Der einstellbare Parameter der Lernrate dient dabei als Schrittweite jedes Optimierungsvorgangs. Praktische Hinweise zur Parameterwahl und der Herangehensweise beim Training von KNN wurden unter anderem von Ng [101] zusammengestellt.

Tiefe KNN sind anfällig für das Problem der Überanpassung, bei dem das Modell zu spezifisch auf die vorliegenden Trainingsdaten optimiert wird und schlecht generalisiert. Abhilfe schafft, neben der oben beschriebenen Datenerweiterung, eine Regularisierung, die z. B. durch *Batch*-Normalisierung [54] oder *Dropout* [128] erreicht werden kann. Beim *Dropout* wird im Training ein vorgegebener Prozentsatz der Verbindungen zwischen Neuronen zufällig nicht berücksichtigt, wodurch die übrigen Verbindungen gestärkt und robuste, redundante Strukturen gefördert werden. Der Prozentsatz wird für jede Schicht separat definiert. Im Falle der *Batch*-Normalisierung wird jeder *Batch* anhand seines Mittelwertes und seiner Standardabweichung normiert. Dies ermöglicht deutlich höhere Lernraten, ein schnelleres Training und eine geringere Abhängigkeit von der anfänglichen Initialisierung der Modellgewichte [54].

Lernbasierte Ansätze werden anhand der verwendeten Trainingsdaten in die drei Kategorien überwacht, halbüberwacht und unüberwacht (englisch *supervised*, *semi-supervised* und *unsupervised*) eingeteilt. Im überwachten Fall liegen während des Trainings sowohl die Eingangsdaten als auch die gewünschten, dazu passenden Ziel- bzw. Ausgangsdaten, häufig *Labels* genannt, vor. Dagegen sind die Zieldaten im unüberwachten Fall nicht bekannt, sodass das zu trainierende Modell die Eingangsdaten aufgrund von Ähnlichkeiten in neue Klassen gruppiert. Beim halbüberwachten Lernen sind sowohl Trainingsdaten mit als auch ohne *Labels* vorhanden, weshalb es Eigenschaften von überwachtem und unüberwachtem Lernen vereinigt [50]. Je mehr *Labels* bekannt sind, desto genauer kann das Modell in der Regel die gewünschten Zusammenhänge lernen. Die Generierung der *Labels* ist aber meist sehr zeitaufwendig.

3 Bewertung der Qualität polyphoner Musikaufnahmen

Vor der digitalen Bearbeitung von Musiksignalen müssen diese zunächst aufgenommen werden. Je nach Aufnahmeszenario, verwendeter Hardware oder zusätzlicher Störsignale schwankt die dabei erzielbare Aufnahmegüte sehr stark. Insbesondere im Falle von Amateuraufnahmen, welche z. B. durch die ständige Verfügbarkeit von Aufnahmemikrofonen in Smartphones weit verbreitet sind, ist eine merklich verringerte Aufnahmequalität gegenüber professionellen Studioaufnahmen zu erwarten. Die reduzierte Qualität beeinträchtigt in den meisten Fällen die nachgeschaltete Signalverarbeitung, woraus schlechtere Ergebnisse für darauf aufbauende MIR-Aufgaben resultieren.

Um diese negativen Auswirkungen zu minimieren, können gezielte Algorithmen zur Signalverbesserung, beispielsweise eine Rauschunterdrückung, ein Herausfiltern von einzelnen Störfrequenzen oder eine Dämpfung definierter gestörter Zeitabschnitte, vor der eigentlich gewünschten Signalverarbeitungsaufgabe durchgeführt werden. Dafür ist eine genaue Bewertung der Aufnahmequalität des zu analysierenden Musikstücks notwendig, die neben der allgemeinen Qualität auch verschiedene Störungsklassen betrachtet. Abhängig von den geschätzten Anteilen jeder Störungsklasse kann dann entschieden werden, welcher Algorithmus zur Signalvorverarbeitung sinnvollerweise eingesetzt wird. Alternativ können in einer Anwendung mit Nutzerinteraktion Hinweise zur Aufnahmeverbesserung direkt an den Nutzer gegeben werden, die anhand der Störungsanteile in der bisherigen Aufnahme identifiziert wurden. Die direkte Rückmeldung zur Aufnahmequalität hilft dem Nutzer zudem bei der Einschätzung der erzielbaren MIR-Ergebnisse.

Bisherige Ansätze zur Bewertung von Audioaufnahmen werden in Abschnitt 3.1 diskutiert. In dieser Arbeit wird die Aufnahmequalität von Musiksignalen anhand von charakteristischen Parametern für die drei

Klassen Hintergrundrauschen, kurze Störgeräusche und Raumakustik bewertet. Dabei werden die Parameter unabhängig voneinander durch kleine neuronale Netze geschätzt, sodass ihr Einfluss getrennt voneinander betrachtet werden kann. Die genaue Umsetzung sowie Ergebnisse der Schätzungen jeder Klasse, welche hauptsächlich für Klaviersignale untersucht werden, sind in den Abschnitten 3.2.1, 3.2.2 und 3.2.3 ausgeführt. Anschließend werden die definierten Qualitätsparameter in Abschnitt 3.3 auf das MIR-Beispiel der Klaviertranskription angewandt. Der Zusammenhang zwischen den Transkriptionsergebnissen und den geschätzten Qualitätsparametern wird für unterschiedliche Aufnahmequalitäten analysiert. Damit lässt sich die erreichte Transkriptionsgüte anhand der geschätzten Parameter zur Aufnahmequalität einordnen.

Die in diesem Kapitel vorgestellten Inhalte zur Qualitätsbewertung polyphoner Klavieraufnahmen und ihre Anwendung in der Transkription wurden bereits in [A7] veröffentlicht.

3.1 Qualitätsbewertung von Audiosignalen

Digitale Audiosignale dienen häufig der Unterhaltung oder Kommunikation von Menschen, weshalb menschliche Testpersonen die Signalqualität von Audioaufnahmen häufig am besten beurteilen können. Um allgemeine Qualitätsaussagen zu treffen, werden Hörtests durchgeführt, in denen Testpersonen durch Zahlenwerte oder eine Reihenfolge der Audiobeispiele die subjektiv empfundene Qualität ausdrücken [81]. Diese subjektiven Qualitätsbewertungen sind sehr zeitaufwendig und kostenintensiv, weshalb objektive Qualitätsmetriken entwickelt wurden, welche die subjektiven Bewertungsergebnisse möglichst gut nachbilden. Ein solches objektives Qualitätsmaß für Sprachsignale in der Telekommunikation stellt die *Perceptual Evaluation of Speech Quality* [110] dar. Sie weist einen an die menschliche Qualitätsbewertung mithilfe von Zahlen zwischen 1 und 5 angepassten Wertebereich zwischen –0,5 und 4,5 auf und ist für diverse Telefonnetze und Codierungen gültig. Ihre Berechnung erfolgt über die Gewichtung von symmetrischen und asymmetrischen Störanteilen, die aus der Differenz der Lautstärkespektren von Originalsignal und übertragenem Signal berechnet werden.

Für Signalverarbeitungsalgorithmen ist eine auf menschlicher Wahrnehmung basierende Qualitätsbewertung im Allgemeinen weniger gut geeignet, da sie subjektive Bewertungseinflüsse enthält. Objektiver und damit besser geeignet sind mathematische Qualitätsmetriken wie das Signal-Rausch-Verhältnis (SNR). Darüber hinaus lässt sich das SNR für bestimmte Arten von Audiosignalen auch bei unbekanntem Originalsignal schätzen. Im Falle von Sprachsignalen modelliert der Algorithmus des *National Institute of Standards and Technology* (NIST) das Rauschen mithilfe eines sequenziellen Gauß'schen Mischungsansatzes [99]. Das SNR wird anschließend aus den Amplitudendichten des modellierten Rauschens und des Gesamtsignals berechnet. Ein zweiter Ansatz, der sich die statistischen Charakteristiken eines Sprachsignals zunutze macht, ist die *Waveform Amplitude Distribution Analysis* (WADA) [64]. Sie nimmt für das Rauschen ebenfalls eine Gauß-Verteilung und für das ungestörte Sprachsignal eine Gamma-Verteilung an. Die Ergebnisse dieser beiden auf klassischen Methoden basierenden Ansätze wurden von der SNR-Schätzung nach Papadopoulos et al. [103] übertroffen, die das SNR kurzer Sprachsignale durch ein neuronales Netz mit vier verborgenen Schichten aus je 1024 Neuronen schätzt. Dabei werden am Netzeingang 712 vorverarbeitete Merkmale eingespeist, die sowohl Energieverhältnisse als auch spezielle Rauschvektoren beinhalten. Dadurch ist die Schätzung unabhängig von der im Signal auftretenden Rauschart. Detailliertere Ausführungen zur subjektiven und objektiven Qualitätsbewertung von Sprachsignalen sind im Überblick von Loizou [81] zu finden.

Im Gegensatz zu Sprachsignalen besitzen Musiksignale längere Passagen definierter Frequenzen und nur selten komplette Pausen, sodass z. B. die vom NIST- und WADA-Algorithmus angenommene Signalcharakteristik nicht zutrifft und diese Ansätze somit für die SNR-Schätzung von Musiksignalen ungeeignet sind. Andere Ansätze dafür sind in der Literatur bisher nicht bekannt. Oft wird eher das Ziel eines für die menschliche Wahrnehmung optimalen Klangerlebnisses verfolgt. Dafür sind subjektive Qualitätsbewertungen, wie beispielsweise für komprimierte und in der Lautstärke angepasste Musiksignale [15], unerlässlich, um eine direkte Rückmeldung potenzieller Nutzer zu erhalten.

Für nachfolgende Algorithmen zur Lösung von MIR-Aufgaben spielt die subjektive Einschätzung dagegen eine untergeordnete Rolle. Deutlich

wichtiger ist die Auswirkung von Signalveränderungen auf nachfolgende MIR-Aufgaben. Dazu haben Mauch und Ewert [87] 14 kontrollierbare Kategorien typischer Beeinträchtigungen von Musiksignalen wie Signalüberlagerung, Hoch- und Tiefpassfilterung, Kompression oder Amplitudenbegrenzung definiert. Die Auswirkung dieser Qualitätsreduktionen wurden für MIR-Algorithmen zur Audio-Identifikation, Takterkennung, Akkorddetektion sowie Ausrichtung von Partitur und Audiosignal analysiert. Dabei konnte keine allgemeine Beziehung zwischen Qualitätsverlust und Güte der MIR-Algorithmen festgestellt werden. Die Ergebnisse einzelner Methoden werden aber stark von bestimmten Kategorien beeinflusst, weshalb Tests mit beeinträchtigten Musiksignalen zur Entwicklung von robusten Verfahren sinnvoll sind.

Im Falle von datengetriebenen Ansätzen kann Robustheit durch einen diversen Datensatz erreicht werden, der verlustbehaftete Signale unterschiedlicher Kategorien enthält. Mithilfe eines solchen Datensatzes konnten beispielsweise mehrere Ansätze zur Erkennung akustischer Ereignisse verbessert werden [122]. Darüber hinaus kann das Hinzufügen von weißem Rauschen oder die Komprimierung des Audiosignals zu einer Robustheit gegenüber gezielt eingefügten Störungen in den Eingangsdaten beitragen [133]. Die Auswirkung der Audiokomprimierung wurde für die MIR-Anwendungen der inhaltsbasierten Musiksuche [41] und der Akkorderkennung [138] genauer untersucht. Während die Akkorderkennung durch eine Komprimierung nicht stark beeinträchtigt wird, machen sich unterschiedliche Bitraten in den Ergebnissen der inhaltsbasierten Musiksuche bemerkbar. Diese Einflüsse können durch Normalisierung der verwendeten Merkmale reduziert werden.

Neben den Qualitätseigenschaften des reinen Signals müssen im Falle von Audioaufnahmen häufig auch die Umgebungsbedingungen der Aufnahme berücksichtigt werden, da sie, wie in Abschnitt 2.1.1 ausgeführt, einen großen Einfluss auf das am Mikrofon ankommende Signal haben. In den meisten Fällen werden dazu die Nachhallzeit t_{RT} und teilweise auch die Anfangsnachhallzeit t_{EDT} der Raumimpulsantwort geschätzt. Kendrick et al. [61] beschreibt ein Verfahren zur Schätzung beider Zeiten auf Basis eines vorab bekannten Musiksignals. Dazu ist eine Vorverarbeitung notwendig, die das Signal zunächst in definierte Frequenzbänder aufteilt, normalisiert und dann die Einhüllende der

Spektren für jedes Band berechnet. Ein KNN schätzt daraus anschließend die zugehörigen RIR-Parameter. Aufgrund ihrer Charakteristik mit immer wieder auftretenden Pausen eignen sich Sprachsignale deutlich besser für eine Schätzung dieser Parameter als Musiksignale [62]. Nur in niedrigen Frequenzbereichen haben Musiksignale teilweise leichte Vorteile, da die Sprachanregung dort zu wenig Energie enthält.

Für den Fall, dass die ursprünglichen Audiosignale unbekannt sind, wurden bisher nur Verfahren zur Schätzung der Nachhallzeit in Sprachaufnahmen entwickelt. Ein Ansatz hierzu ist die statistische Modellierung des Schallabfalls mit anschließender iterativer Optimierung [82]. Der Schallabfall kann auch aus der STFT des Eingangssignals berechnet werden, wodurch die Schätzung der Nachhallzeit schnell und robust gegen additives Rauschen ist [26]. Für den Einsatz in mobilen Anwendungen wie digitalen Hörgeräten haben Diether et al. [22] einen echtzeitfähigen Algorithmus zur Schätzung der Nachhallzeit entwickelt, der eine Aufteilung in Frequenzbänder nutzt. Alle diese Ansätze zur „blinden“ Schätzung von RIR-Parametern sind sehr auf die Charakteristik von Sprachsignalen angepasst und deshalb nicht für Musiksignale einsetzbar.

3.2 Schätzung charakteristischer Qualitätsparameter

Um die Aufnahmequalität von polyphonen Musiksignalen objektiv bewerten zu können, werden in dieser Arbeit charakteristische Parameter der drei Haupteinflussfaktoren Hintergrundrauschen, kurze Störgeräusche und Nachhall geschätzt. Diese drei Einflussklassen sind die häufigsten Ursachen für eine reduzierte Qualität von Amateuraufnahmen. Andere Einflüsse wie Normalisierung und Komprimierung spielen eine untergeordnete Rolle und beeinträchtigen manche MIR-Aufgaben kaum (s. Abschnitt 3.1), sodass sie hier vernachlässigt werden. Aus den definierten Haupteinflussfaktoren resultieren die Parameter konstantes SNR, zeitabhängiges SNR sowie die Nachhallzeiten t_{RT} und t_{EDT}. Sie betrachten sehr unterschiedliche Eigenschaften des aufgenommenen Signals, weshalb sie durch voneinander unabhängige Ansätze geschätzt werden. Das Schema dieser separaten Schätzung ist mit den zugehörigen Parametern in Abbildung 3.1 dargestellt.

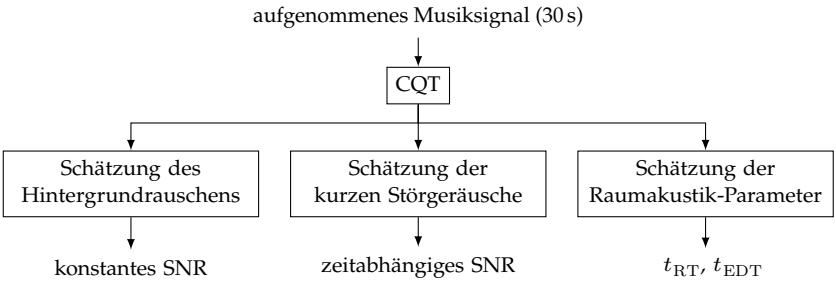


Abbildung 3.1 Schema der Schätzung aller betrachteten Qualitätsparameter.

Alle drei Ansätze werden mithilfe künstlicher neuronaler Netze umgesetzt, da KNN sowohl im Falle der SNR-Schätzung von Sprachsignalen die besten Ergebnisse gezeigt haben [103] als auch zur Schätzung der RIR-Parameter mit bekanntem Musiksignal erfolgreich verwendet wurden [61]. In Vorversuchen bestätigte sich die mangelnde Übertragbarkeit bestehender Verfahren für Sprachsignale auf Musiksignale, weshalb die KNN zur Parameterschätzung in dieser Arbeit neu designt werden. Dabei liegt der Fokus auf kleinen, aber leistungsstarken Netzen, sodass eine schnelle und mit geringem Aufwand realisierbare Bewertung der Aufnahmequalität von Musiksignalen möglich ist.

Der erste Schritt zu kleinen KNN ist die Verwendung weniger, aussagekräftiger Merkmale am Eingang. Daher wird jede zu analysierende Musikaufnahme in ein Mono-Signal überführt und daraus der Betrag $\left|X_{\mathrm{CQT}}[m,k]\right|$ der in Abschnitt 2.2.2 beschriebenen CQT berechnet. Sie enthält spektrale Informationen in 84 Frequenzbins, die sieben Oktaven zwischen den Noten C_1 (32,7 Hz) und c^5 (4186 Hz) umfassen. Aus der vorgegebenen Länge jedes Eingangssignals von 30 s folgt bei einer *Hop Size* von 512 und der in der Audiosignalverarbeitung üblichen Abtastfrequenz 22 050 Hz eine zeitliche Dimension von 1292 Werten.

Alle in diesem Kapitel vorgestellten KNN werden mithilfe des Adam-Optimierers [66] und der quadratischen Kostenfunktion MSE (s. Gleichung (2.27)) über 50 Epochen trainiert. Jeder *Batch* umfasst dabei eine Größe von 1024 Musiksignalen. Diese Musiksignale können theoretisch von verschiedenen Instrumenten stammen. In dieser Arbeit liegt der Fokus allerdings auf Klaviersignalen, da diese in großer Zahl verfügbar

sind und in mehreren Anwendungen wie der Transkription eine große Rolle spielen. Ein Beispiel solcher großen Datensätze ist der hier verwendete MAPS-Datensatz [27], welcher Aufnahmen von 270 Klavierstücken sowie einer Vielzahl von Einzelnoten und Einzelakkorden beinhaltet. Für die Schätzung der Aufnahmequalität werden ausschließlich die 270 Klavierstücke verwendet, weil sie realitätsnahe Aufnahmeszenarien repräsentieren. Sie setzen sich aus 210 synthetisch generierten Stücken und 60 realen Klavieraufnahmen zusammen. Die ersten 30 s aller synthetischen Stücke werden als Trainingsdaten und die ersten 30 s aller realen Aufnahmen als Testdaten verwendet.

Obwohl die KNN im Training hinsichtlich des MSE optimiert werden, wird zur Analyse der Ergebnisse in den folgenden Abschnitten der mittlere absolute Fehler (englisch *mean absolute error*, MAE) verwendet. Er berechnet sich über

$$\mathrm{MAE} = \frac{1}{L} \sum_{i=1}^{L} \left| y_i - \hat{y}_i \right| \tag{3.1}$$

und gibt eine mittlere Abweichung der Schätzung $\hat{y}_i$ vom Zielwert y_i an, weshalb er anschaulicher als der auf dem Fehlerquadrat beruhende MSE ist. Weitere Evaluationen der vorgestellten Qualitätsparameterschätzung werden in Abschnitt 3.3 im Hinblick auf ihre Anwendbarkeit für die Klaviertranskription durchgeführt.

3.2.1 Hintergrundrauschen

Der Einfluss von Hintergrundrauschen auf die Aufnahmequalität wird durch einen SNR-Wert für jedes 30 s lange Musiksignal bewertet, der mithilfe eines KNN geschätzt wird. Über diesen Zeitraum wird die Rauschcharakteristik als näherungsweise konstant angenommen. Vor der Schätzung werden die Merkmale der CQT des Musiksignals in einem zusätzlichen Vorverarbeitungsschritt weiter komprimiert, indem für jedes der 84 Frequenzbänder der Mittelwert und die Varianz über alle 1292 Zeitbins berechnet werden. Somit muss das KNN zur Schätzung des Hintergrundrauschens nur 168 Eingangswerte verarbeiten, was sowohl den Ressourcenbedarf als auch die Laufzeit drastisch verringert.

Aufgrund der Vorverarbeitung reicht ein kleines Netz aus vier voll verbundenen Schichten zur Schätzung des SNR-Wertes aus, dessen Ar-

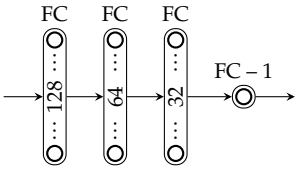


Abbildung 3.2 Netzarchitektur aus voll verbundenen Schichten (FC) zur SNR-Schätzung bei Hintergrundrauschen.

chitektur in Abbildung 3.2 dargestellt ist. Die Anzahl der Merkmale reduziert sich schrittweise bis zur Ausgangsschicht, die nur aus einem Neuron mit linearer Aktivierungsfunktion besteht. Alle verborgenen Schichten nutzen die ReLU-Funktion als Aktivierung und während des Trainings *Dropout* mit einer Rate von 40 %, um Überanpassung zu verhindern. Insgesamt besitzt das beschriebene KNN nur 32 001 Parameter und ist damit z. B. deutlich kleiner als das mehr als zehnmal so große KNN von Papadopoulos et al. [103] zur SNR-Schätzung bei Sprachsignalen.

Als Hintergrundrauschen werden drei generische Rauschtypen und eine Aufnahme von hochfrequenten Funkgeräuschen [57] berücksichtigt. Die generischen Typen sind weißes Rauschen sowie einfach und zweifach tiefpassgefiltertes weißes Rauschen, die häufig als pinkes und braunes Rauschen bezeichnet werden. Pinkes Rauschen besitzt eine konstante Energiedichte pro Oktave und braunes Rauschen ist ein Beispiel für tieffrequentes Rauschen. Alle vier Rauscharten erfüllen die Annahme der stationären Rauschcharakteristik. Um unterschiedliche Signal-Rausch-Verhältnisse abzubilden, wird jedes erzeugte Rauschsignal gewichtet und anschließend mit einer Musikaufnahme des Trainings- oder Testdatensatzes überlagert. Der Gewichtungsfaktor wird jeweils so gewählt, dass der resultierende SNR-Wert, welcher über die mittleren Signalenergien von Musikstück und Rauschen berechnet wird, einem ganzzahligen Vielfachen von 2,5 dB im Intervall [–5 dB, 20 dB] entspricht. Dabei wird angenommen, dass die originalen Klavieraufnahmen des MAPS-Datensatzes nur sehr geringes Rauschen aufweisen, das vernachlässigt werden kann. Im resultierenden Trainingsdatensatz ist jede Kombination aus MAPS-Musikstück, Rauschtyp und zulässiger SNR-Stufe fünfmal enthal-

ten, im Testdatensatz nur einmal. Daraus ergeben sich 46 200 Trainings- und 2640 Testbeispiele für die vier Rauscharten und elf SNR-Stufen.

Neben den reinen Klavieraufnahmen des MAPS-Datensatzes werden synthetisch erzeugte Musikstücke anderer Instrumente im Training integriert, um die Menge der Trainingsdaten anzuheben und einen diverseren Datensatz zu generieren. Dazu werden die MAPS-Musikstücke anhand ihrer MIDI-Dateien durch neun verschiedene Klangprofile (KP) des *General MIDI 1 Sound Set* [92] mithilfe des Programms *pretty_midi* [106] neu synthetisiert. Folglich weisen diese synthetischen Daten kein aufnahmebedingtes Rauschen auf und sie können wie oben beschrieben durch Überlagerung von Hintergrundrauschen auf die vorgegebenen SNR-Stufen gebracht werden. Aus der Kombination aller MAPS-Musikausschnitte, Rauschtypen, SNR-Stufen und Klangprofile resultieren 83 160 synthetisch generierte Trainingsdaten sowie 23 760 synthetische Testdaten. Durch die Wahl der synthetischen Klangprofile Konzertflügel (KP_1), Kirchenorgel (KP_{20}), akustische Gitarre (KP_{25}), akustische Bassgitarre (KP_{33}), Viola (KP_{42}), Trompete (KP_{57}), Tenorsaxofon (KP_{67}), Flöte (KP_{74}) und Banjo (KP_{106}) wird ein breites Instrumentenspektrum abgedeckt. Die Zahlen repräsentieren dabei die entsprechenden MIDI-Nummern der Instrumente.

Das KNN wird zunächst nur mit dem ersten Trainingsdatensatz aus 46 200 verrauschten MAPS-Aufnahmen trainiert, damit keine fehlerhaften Bezüge aus den synthetischen Klangprofilen gelernt werden. Anschließend folgt ein zweites Training mit dem Trainingsdatensatz der synthetisch generierten Instrumentensignale, wodurch die Gewichte des KNN weiter optimiert werden. Durch diese zweistufige Vorgehensweise, dem damit verbundenen längeren Training sowie die diverseren Trainingsbeispiele verbessern sich die Ergebnisse der SNR-Schätzung sowohl für die auf realen Aufnahmen als auch für die auf synthetisch generierten Musikstücken basierenden Testdaten um etwa 1 dB. Die genauen Ergebnisse der Schätzung nach zweistufigem Training sind in Tabelle 3.1 für die vier analysierten Arten von Hintergrundrauschen aufgeführt, wobei die auf real aufgenommenen und synthetisierten Musiksignalen basierenden Testdatensätze separat betrachtet werden.

Für beide Testdatensätze ist die SNR-Schätzung bei Überlagerung von braunem Rauschen am genauesten. Im Vergleich zu den Rauschtypen mit

Tabelle 3.1 MAE (in dB) der SNR-Schätzung für die auf realen Aufnahmen und synthetisch generierten Musikstücken basierenden Testdatensätze bei verschiedenen Arten von Hintergrundrauschen.

	weiß	pink	braun	hochfreq.	∅
Reale Aufnahmen	0,95	0,95	0,85	1,09	0,96
Synth. Musikstücke	1,80	1,67	1,54	1,77	1,69

den schlechtesten Ergebnissen ist der MAE bei braunem Rauschen aber nur um maximal 0,26 dB niedriger, sodass zwischen den Rauscharten eine vergleichbare Güte festgestellt werden kann. Größere Unterschiede treten zwischen den Testdatensätzen auf. Obwohl das KNN im zweiten Schritt mit den synthetisierten Trainingsstücken trainiert wurde, liefert es für reale Aufnahmen mit überlagertem Hintergrundrauschen im Durchschnitt um ca. 0,7 dB bessere Ergebnisse, wodurch sich der MAE fast halbiert. Dies bestätigt die Annahme von wenig Rauschen in den originalen Klavieraufnahmen, da die SNR-Werte der synthetisch generierten, zunächst rauschfreien Musikstücke nach dem zweiten Training ansonsten deutlich besser geschätzt werden müssten.

Eine Erklärung für die Differenz zwischen den Ergebnissen der beiden Testdatensätze ist die Integration sehr unterschiedlicher Instrumentencharakteristiken bei der Erzeugung der synthetisierten Musikstücke. Spezielle Klangprofile wie Kirchenorgel oder akustische Bassgitarre erschweren durch einen breiten Registerumfang oder einen deutlich abweichenden Frequenzbereich die Schätzung des Hintergrundrauschens. Diese Einschätzung wird durch den durchschnittlichen MAE und die mittlere Standardabweichung (STD) über alle vier Rauscharten untermauert, die in Tabelle 3.2 für die realen Aufnahmen sowie jedes Klangprofil angegeben sind. Beide Metriken sind im Falle von Kirchenorgel (KP_{20}) und Bassgitarre (KP_{33}) deutlich höher als für die übrigen Instrumente.

Mit Ausnahme der Kirchenorgel liegen alle durchschnittlichen MAE innerhalb von einer SNR-Stufe von 2,5 dB. Darüber hinaus beträgt die durch die STD repräsentierte Streuung der geschätzten SNR-Werte in den meisten Fällen deutlich weniger als 2,5 dB, nur die Werte für Kirchenorgel und akustische Bassgitarre liegen knapp darüber. Daraus folgt,

Tabelle 3.2 Durchschnittliche MAE und STD der SNR-Schätzung bei Hintergrundrauschen aller vier betrachteten Rauscharten, aufgeschlüsselt nach den Klangprofilen Konzertflügel (KP_1), Kirchenorgel (KP_{20}), akust. Gitarre (KP_{25}), akust. Bassgitarre (KP_{33}), Viola (KP_{42}), Trompete (KP_{57}), Tenorsaxofon (KP_{67}), Flöte (KP_{74}) und Banjo (KP_{106}).

	Real	KP_1	KP_{20}	KP_{25}	KP_{33}
MAE (in dB)	0,96	1,19	3,06	1,15	2,14
STD (in dB)	0,99	1,17	2,64	1,16	2,59

	KP_{42}	KP_{57}	KP_{67}	KP_{74}	KP_{106}
MAE (in dB)	1,41	1,65	1,75	1,12	1,78
STD (in dB)	1,34	1,72	1,80	1,23	1,57

dass die vorgestellte SNR-Schätzung für Musiksignale mit Hintergrundrauschen im Allgemeinen zuverlässig funktioniert. Das gilt sowohl für die mit Rauschen überlagerten realen Klavieraufnahmen, für welche die genauesten Schätzergebnisse erzielt werden können, als auch für verrauschte Musikstücke anderer Instrumente.

Die Hinzunahme von weiteren Trainingsbeispielen durch synthetische Klangprofile erreicht zwar eine leichte Verbesserung der Schätzgenauigkeit, ist aber zeitaufwändig und benötigt viele Ressourcen. Des Weiteren stellt die direkte Übertragung von polyphoner Klaviermusik auf andere Instrumente ein teilweise unrealistisches Szenario dar, weshalb diese Form der Datenerweiterung in den folgenden Ansätzen zur Qualitätsbewertung nicht weiter genutzt wird.

3.2.2 Kurze Störgeräusche

Abgesehen vom in Abschnitt 3.2.1 untersuchten Hintergrundrauschen können in Audioaufnahmen Störgeräusche auftreten, deren Einfluss auf eine kürzere Zeit begrenzt ist. In diesem Zeitraum sinkt das resultierende SNR auf einen niedrigeren Wert ab. Folglich ist die Schätzung eines zeitabhängigen Signal-Rausch-Verhältnisses notwendig, um die Aufnahmequalität in Abhängigkeit der Zeit bewerten zu können. Diese Zeitabhängigkeit wird durch die separate Schätzung des SNR für

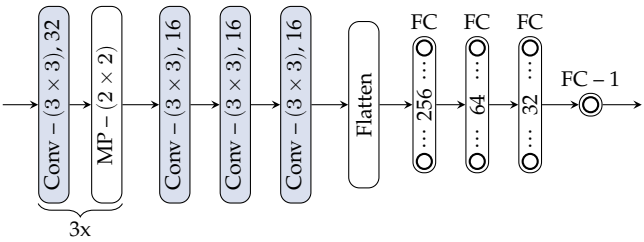


Abbildung 3.3 Schema der Netzarchitektur zur SNR-Schätzung von kurzen Zeitabschnitten. Der erste Teil besteht aus Faltungs- (Conv) und *Max-Pooling*-Schichten (MP), der zweite Teil aus voll verbundenen Schichten (FC).

sehr kurze, aufeinanderfolgende Signalabschnitte erreicht. In dieser Arbeit werden Abschnitte von 2 s Länge betrachtet, die 87 Zeitbins in der CQT entsprechen. Da kurze Störgeräusche in der Regel keine stationäre Störcharakteristik aufweisen, reicht die Betrachtung von Mittelwert und Standardabweichung aus Abschnitt 3.2.1 nicht aus. Daher wird die gesamte CQT des zu analysierenden Signalabschnitts, welche die Dimension $(87, 84)$ besitzt, am Eingang des KNN eingespeist.

Um die Dimension der zweidimensionalen Eingangsdaten zunächst zu reduzieren und komplexe Merkmale zu extrahieren, besteht der erste Teil des KNN zur SNR-Schätzung kurzer Zeitabschnitte aus zweidimensionalen Faltungs- und *Max-Pooling*-Schichten. Anschließend wird der SNR-Wert im zweiten Teil mithilfe von voll verbundenen Schichten geschätzt, die nahezu dem KNN zur SNR-Schätzung bei Hintergrundrauschen entsprechen. Daraus ergibt sich eine klassische CNN-Architektur, deren genaue Umsetzung in Abbildung 3.3 dargestellt ist. Die *Flatten*-Schicht ist dabei zur Verknüpfung der beiden KNN-Teile notwendig, da sie das zweidimensionale Faltungsergebnis in einen eindimensionalen Vektor transformiert. Sowohl die Faltungsschichten als auch die verborgenen voll verbundenen Schichten nutzen die ReLU-Aktivierungsfunktion. Wie schon im Fall von Hintergrundrauschen beinhalten die verborgenen voll verbundenen Schichten zusätzlich *Dropout* mit einer Rate von 40 %. Insgesamt enthält das beschriebene KNN zur SNR-Schätzung kurzer Zeitabschnitte 71 473 Parameter, wodurch auch für diesen Fall des zeitabhängigen SNR eine effiziente Qualitätsbewertung möglich ist.

Tabelle 3.3 MAE und STD der SNR-Schätzung bei Störgeräuschen der Klassen Klimaanlage (a), Autohupen (b), Hundebellen (c), Kantine (d), Fabriklärm (e) und Kinderspielplatz (f).

	(a)	(b)	(c)	(d)	(e)	(f)	∅
MAE (in dB)	2,80	2,75	2,23	1,87	1,75	2,40	2,30
STD (in dB)	3,05	2,51	2,48	2,04	1,97	2,68	2,46

Zeitlich beschränkte Störgeräusche können mit sehr unterschiedlicher Charakteristik auftreten, weshalb im Trainings- und Testdatensatz diverse Arten von Störgeräuschaufnahmen enthalten sind. Zum einen werden Geräusche mit Grundpegel und einzelnen Störimpulsen wie Kantinen- oder Fabriklärm [57] betrachtet, zum anderen werden aus dem Urban-Sound-Datensatz [114] impulsartige Störer wie Hundebellen oder Autohupen sowie monotone Geräusche wie das einer Klimaanlage verwendet. Darüber hinaus werden im Testdatensatz Audioaufnahmen eines Spielplatzes mit spielenden Kindern [114] als eine im Training unbekannte Geräuschklasse berücksichtigt. Zur Datensatzgenerierung werden aus jedem 30 s langen MAPS-Musiksignal zehn kurze, nicht überlappende Abschnitte von 2 s extrahiert und mit zufälligen Ausschnitten der Störgeräuschaufnahmen überlagert. Analog zum Vorgehen in Abschnitt 3.2.1 werden die Geräuschaufnahmen vor der Überlagerung mit einem Faktor gewichtet, sodass sich SNR-Stufen eines ganzzahligen Vielfachen von 2,5 dB im Intervall [−5 dB, 20 dB] ergeben. Dabei wird das SNR anhand der mittleren Signalenergien der beiden 2 s langen Abschnitte von Musik- und Störgeräuschaufnahme berechnet. Für den Trainingsdatensatz wird der Generierungsprozess für alle MAPS-Musikaufnahmen, fünf Geräuschklassen und elf SNR-Stufen viermal durchlaufen, sodass er aus 462 000 Beispielen besteht. Als Testdaten werden für jede Kombination der elf SNR-Stufen und sechs betrachteten Geräuschklassen (fünf im Training bekannte plus Kinderspielplatz) 200 zufällige Musikabschnitte überlagert, sodass der Testdatensatz 13 200 Beispiele umfasst.

Die Güte der SNR-Schätzung für kurze Musikabschnitte von 2 s Länge wird anhand des durchschnittlichen MAE und der mittleren STD bewertet, welche in Tabelle 3.3 für jede betrachtete Geräuschklasse aufgelistet sind. Das SNR kann für das monotone Geräusch der Klimaanlage am

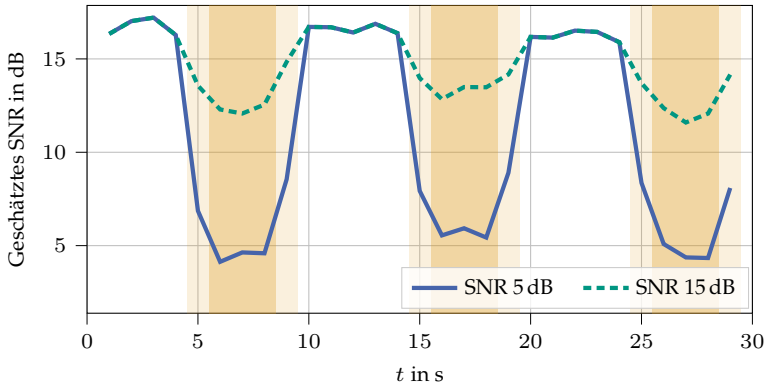


Abbildung 3.4 Zeitlicher Verlauf des durchschnittlich geschätzten SNR-Wertes für zeitweise überlappendes Hundebellen mit zwei verschiedenen Lautstärken. Die komplett durch Bellen überlagerten Musiksignalabschnitte sind in orange, die halb überlagerten in hellorange gekennzeichnet.

schlechtesten geschätzt werden. Da es das einzige monotone Geräusch im Trainingsdatensatz darstellt, kann die kleinere Beispielmenge dieser Geräuschart eine Ursache für die verringerte Güte sein. Bei impulsartigen Störungen ist die Schätzung nur leicht besser, was an der speziellen Charakteristik mit stark zeitabhängiger Signalenergie liegt. Die besten Ergebnisse werden bei Geräuschen mit Grundpegel und einzelnen Störimpulsen erzielt. Sogar die unbekannte Geräuschklasse Kinderspielplatz, die eine ähnliche Charakteristik aufweist, erreicht Metriken im Bereich der Mittelwerte aller Klassen. Mit durchschnittlichen MAE- und STD-Werten von unter 2,5 dB bleibt die mittlere Abweichung und Streuung der SNR-Schätzung innerhalb einer SNR-Stufe, was eine zuverlässige Detektion von Signalabschnitten mit deutlich verringerter Aufnahmequalität ermöglicht.

Die zeitabhängige SNR-Schätzung wird durch die Abbildung 3.4 illustriert, welche die über mehrere Teststücke von 30 s gemittelten Schätzergebnisse bei dreifacher kurzer Störung durch Hundebellen zeigt. Um den zeitlichen Verlauf genauer abzubilden, werden Signalabschnitte von 2 s Länge extrahiert, die jeweils 1 s Überlappung zum vorhergehenden und nachfolgenden Abschnitt aufweisen. Deshalb treten in Abbildung 3.4

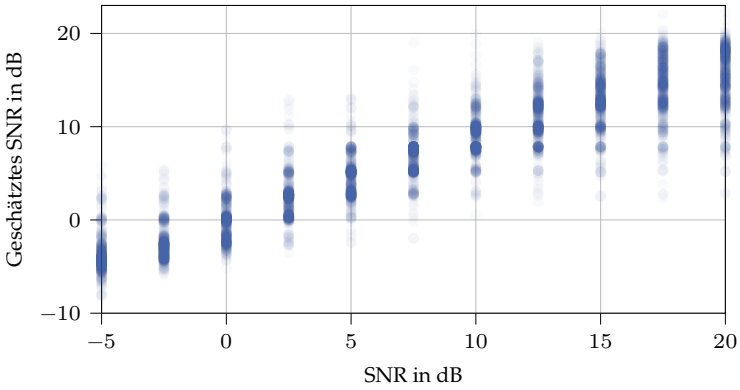


Abbildung 3.5 Verteilung der geschätzten SNR-Werte über alle SNR-Stufen der generierten Testsignale für überlagertes Hundebellen.

die Zeitabschnitte in hellorange auf, deren Musiksignale nur zur Hälfte durch Hundebellen überlagert sind. Das in den Abschnitten mit Hundebellen reduzierte SNR wird durch die Schätzung korrekt durch niedrigere Werte als in den ungestörten Abschnitten erfasst. Für die halb überlagerten Abschnitte werden SNR-Werte zwischen denen der vollständig und nicht gestörten Signalabschnitte geschätzt, was ebenfalls der Erwartung entspricht. Im Falle der ungestörten Abschnitte wird zwar das höchste SNR im gesamten Musikstück geschätzt, allerdings liegen die Werte mit etwa 17 dB unter dem maximal im Trainingsdatensatz vorhandenen SNR von 20 dB. Ein Grund könnten sehr leise Störgeräusche in Teilen der originalen MAPS-Aufnahmen sein, sodass manche „ungestörten“ Musiksignale schon SNR-Werte von unter 20 dB besitzen. Darüber hinaus ist die SNR-Schätzung von unbekannten Musikaufnahmen bei hohen SNR-Werten sehr schwierig, da überlagerte Störgeräusche nur mit sehr geringer Signalenergie auftreten und deshalb teilweise nur bei bekanntem Originalsignal detektiert werden können. Diese Effekte führen im Falle leiser Störgeräusche bzw. hoher SNR-Werte für alle Geräuschklassen zu einer größeren Varianz der geschätzten SNR-Werte. Exemplarisch ist dies in der Verteilung der geschätzten Werte bei Störung durch Hundebellen in Abbildung 3.5 dargestellt. Im Allgemeinen können die mit Störsignalen überlagerten Zeitabschnitte aber anhand ihrer

kleineren SNR-Schätzwerte erkannt werden (s. Abbildung 3.4), weshalb die zeitabhängige SNR-Schätzung die geforderten Ziele erfüllt.

3.2.3 Raumakustik-Parameter

Die SNR-Schätzungen der vorhergehenden Abschnitte bewerten das Eingangssignal hinsichtlich vorhandener Störgeräusche, aber ohne Berücksichtigung der Umgebungsbedingungen während der Aufnahme. Da diese jedoch oft einen hohen Einfluss auf das aufgenommene Signal haben, werden im dritten Teil der Qualitätsbewertung die Nachhall- und Anfangsnachhallzeit t_{RT} bzw. t_{EDT} geschätzt. Als charakteristische Parameter der Raumimpulsantwort (RIR) ermöglichen sie eine einfache Beschreibung der Raumakustik. In dieser Arbeit werden sowohl die Richtungsabhängigkeit der Schallabstrahlung als auch die Frequenzabhängigkeit der Nachhallzeit vernachlässigt, weil monophone Aufnahmen bewertet werden sollen und die Klavieraufnahmen häufig nur Töne im mittleren Frequenzbereich beinhalten, welcher bei Einzahlwerten für die Nachhallzeit typischerweise betrachtet wird [149].

Wie in Abschnitt 2.1.1 ausgeführt, beschreiben die Nachhall- und Anfangsnachhallzeit das zeitliche Abklingen der Signalenergie. Für die Eingangsdaten der Schätzung steht somit die zeitliche Änderung der CQT jedes 30 s langen Musiksignals im Vordergrund, weshalb ein zusätzlicher Vorverarbeitungsschritt die Ableitung der logarithmischen CQT-Werte nach der Zeit berechnet, die häufig auch zur Anschlagsdetektion verwendet wird [89]. Durch die diskreten Zeitbins der CQT repräsentiert jeder Wert dieser Vorverarbeitung die Stärke des Abklingens eines Zeitschritts und eines Frequenzbins. Hinsichtlich kleiner, effizienter Netze werden jeweils vier der aus der Vorverarbeitung resultierenden Zeitreihen von nebeneinander liegenden Frequenzbins gemittelt. Diese Reduktion der Eingangsdaten hebt die zeitlichen Zusammenhänge weiter hervor und verringert die Anzahl der Frequenzbins auf 21. Die Dimension der vorverarbeiteten Eingangsdaten zur Schätzung der Nachhallzeiten beträgt folglich für jedes 30 s lange Musiksignal $(1292, 21)$.

Aufgrund der ebenfalls zweidimensionalen Eingangsdaten besitzt das KNN eine ähnliche Architektur wie das KNN zur SNR-Schätzung kurzer Zeitabschnitte. Die genaue Umsetzung dieser klassischen CNN-Architektur ist in Abbildung 3.6 dargestellt. Zunächst führen die Faltungs-

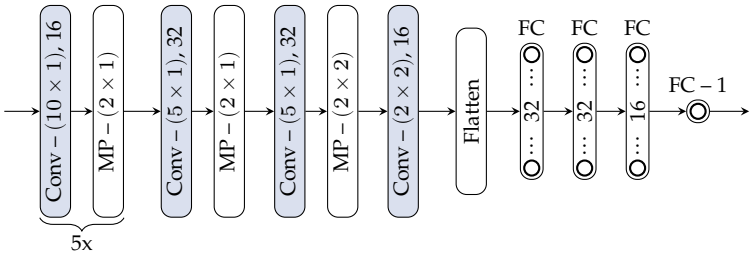


Abbildung 3.6 Schema der Netzarchitektur zur Schätzung der Raumakustik-Parameter t_{RT} und t_{EDT}. Der erste Teil besteht aus Faltungs- (Conv) und *Max-Pooling*-Schichten (MP), der zweite Teil aus voll verbundenen Schichten (FC).

und *Max-Pooling*-Schichten eine Merkmalsextraktion sowie eine Dimensionsreduktion hauptsächlich im Zeitbereich durch. Anschließend wird daraus die Nachhallzeit t_{RT} oder die Anfangsnachhallzeit t_{EDT} mithilfe von voll verbundenen Schichten geschätzt. Zwischen den beiden Teilen wird zur Dimensionsanpassung eine *Flatten*-Schicht verwendet. Sie überführt alle zweidimensionalen Merkmale in einen eindimensionalen Vektor, indem sie alle Datenpunkte nebeneinander aufreiht. Der Fokus auf die Zeitdimension wird auch in der Wahl der Faltungskerne deutlich, die fast alle nur einen Frequenzbin, aber dafür zehn bzw. fünf Zeitbins umfassen. In allen Faltungs- und verborgenen voll verbundenen Schichten wird die ReLU-Aktivierungsfunktion eingesetzt. Das Ausgangsneuron nutzt dagegen die lineare Aktivierung. Analog zu den anderen in diesem Kapitel vorgestellten KNN-Architekturen wird in den verborgenen voll verbundenen Schichten *Dropout* mit einer Rate von 40 % eingesetzt. Daraus ergibt sich für das KNN zur Schätzung der Nachhall- oder Anfangsnachhallzeit eine Architektur mit 35 745 Parametern.

Zur Datengenerierung werden aufgenommene Raumimpulsantworten eines Datensatzes, der 1426 RIRs aus neun Räumen enthält [135], nach Gleichung (2.5) mit den MAPS-Musiksignalen gefaltet. Dabei wird angenommen, dass die Raumakustik der ursprünglichen MAPS-Aufnahmen nur eine untergeordnete Rolle spielt und deswegen vernachlässigt werden kann. Darüber hinaus ändert sich die Raumakustik innerhalb der betrachteten 30 s wenn überhaupt nur marginal, weshalb keine Zeitabhängigkeit der Nachhallzeiten berücksichtigt wird und ein Schätzwert

pro Musikstück ausreicht. Der RIR-Datensatz beinhaltet Raumimpulsantworten unterschiedlicher Positionen von Quelle und Mikrofon in kleineren Räumen wie Büro- oder Hotelzimmern, einem Treppenhaus sowie größeren Konferenz- und Vorlesungsräumen. Damit deckt er viele typische Aufnahmebedingungen hinsichtlich der Raumakustik ab. Zwei Räume (Hotelzimmer R112 und Konferenzraum CR2) werden ausschließlich für den Testdatensatz verwendet, die RIRs der anderen sieben Räume werden zur Trainingsdatenerzeugung eingesetzt. Jede originale MAPS-Aufnahme des Trainingsdatensatzes wird mit 500 zufälligen RIR-Aufnahmen gefaltet, wodurch etwa 92 000 verschiedene Trainingsbeispiele zur Verfügung stehen. Die Generierung der Testdaten erfolgt analog mit 25 zufälligen RIR-Aufnahmen pro MAPS-Teststück, wobei sowohl 1500 Testbeispiele für die beiden unbekannten Testräume als auch für die sieben im Training gelernten Räume erstellt werden.

Aus den RIR-Aufnahmen lassen sich die Wertebereiche der Nachhallzeiten zu $t_{\mathrm{RT}} \in [0{,}4\,\mathrm{s}, 2{,}2\,\mathrm{s}]$ und $t_{\mathrm{EDT}} \in [0{,}2\,\mathrm{s}, 3{,}0\,\mathrm{s}]$ für die Trainingsräume und $t_{\mathrm{RT}} \in [0{,}4\,\mathrm{s}, 2{,}0\,\mathrm{s}]$ sowie $t_{\mathrm{EDT}} \in [0{,}3\,\mathrm{s}, 1{,}5\,\mathrm{s}]$ für die Testräume berechnen. Um die Zeiten besser vergleichen zu können, entsprechen t_{RT} und t_{EDT} den Dauern für einen Energieabfall von 60 dB, obwohl sie, wie in Abschnitt 2.1.1 beschrieben, über kleinere Energieintervalle berechnet werden. Die zunächst berechneten Zeiten $t_{\mathrm{RT}_{20}}$, welche dem Abfall der Schallenergie von –5 dB auf –25 dB entspricht, und $t_{\mathrm{EDT}_{10}}$, die das Abklingen auf –10 dB umfasst, werden deshalb durch Multiplikation mit 3 bzw. 6 auf 60 dB Energieabfall extrapoliert.

Nachhall- und Anfangsnachhallzeit werden in separaten KNN der vorgestellten Architektur geschätzt, wobei sich ihr Trainingsprozess nur hinsichtlich der *Labels* unterscheidet. Die Ergebnisse der Schätzungen von t_{RT} und t_{EDT} sind in Tabelle 3.4 für die Testdaten mit RIRs aus den sieben bekannten Trainings- und den zwei unbekannten Testräumen angegeben. Entgegen der Erwartung, dass die Zeiten der im Training gelernten Räume besser geschätzt werden können, sind die mittleren Fehler beider Parameter für die Testräume etwas kleiner. Ein Grund dafür kann der breitere Wertebereich der Trainingsräume sein, wodurch der Trainingsdatensatz RIRs mit längeren Nachhallzeiten enthält und damit auch ein höheres Risiko für größere absolute Fehler besitzt. Im Vergleich zwischen den beiden Zeiten schneidet die Schätzung der Anfangsnach-

Tabelle 3.4 MAE (in s) der Schätzung von Nachhallzeit t_{RT} und Anfangsnachhallzeit t_{EDT} für die Testdatensätze der sieben Trainings- und zwei Testräume.

	Trainingsräume	Testräume
t_{RT}	0,316	0,288
t_{EDT}	0,224	0,201

hallzeit t_{EDT} deutlich besser ab als die Schätzung der Nachhallzeit t_{RT}. Dies gilt sowohl für die Trainings- als auch die Testräume. Für einige Räume nimmt die Anfangsnachhallzeit dabei kleinere Werte als die Nachhallzeit an, sodass der absolute Fehler bei ähnlicher relativer Abweichung kleiner ist. Beispielsweise ist der niedrigste Werte der Trainingsräume mit $t_{\mathrm{EDT}} = 0{,}2\,\mathrm{s}$ halb so groß wie der der Nachhallzeit. Darüber hinaus wird in der Anfangsnachhallzeit nur die Dauer des Energieabfalls über die ersten 10 dB berücksichtigt, was einfacher zu schätzen sein kann. Insgesamt können bei unbekanntem Klaviersignal aber sowohl t_{RT} als auch t_{EDT} ausreichend gut mithilfe der entwickelten KNN geschätzt werden, da der MAE für Trainingsräume und Testräume maximal im Bereich der kleinsten Nachhallzeiten des RIR-Datensatzes liegt. Während des Trainings tritt dabei auch keine Überanpassung hinsichtlich der verwendeten RIR-Aufnahmen der Trainingsräume auf, weil die Zeiten der unbekannten Testräume sogar etwas besser geschätzt werden.

Einen Überblick über die Güte und Verteilung aller Schätzungen der Nachhallzeit gibt Abbildung 3.7. Für die Anfangsnachhallzeit ergibt sich eine vergleichbare Verteilung. Aufgrund der ähnlichen Ergebnisse von Testdaten mit bekannten und unbekannten Räumen sind alle Datenpunkte der beiden Testdatensätze abgebildet. Somit sind Schätzungen zu neun verschiedenen Räumen enthalten, wobei gut zu erkennen ist, dass in den verwendeten RIRs keine Gleichverteilung der Nachhallzeiten vorliegt. Dies liegt hauptsächlich an den im Datensatz enthaltenen Raumtypen. Die meisten RIRs besitzen Nachhallzeiten um etwa 0,5 s, die vom entwickelten KNN gut geschätzt werden. Größere Nachhallzeiten werden dagegen systematisch unterschätzt, was auf den unbalancierten Datensatz zurückzuführen ist, der mehr Räume mit moderatem Nachhall und damit kleineren Nachhallzeiten enthält. Darüber hinaus spielt bei

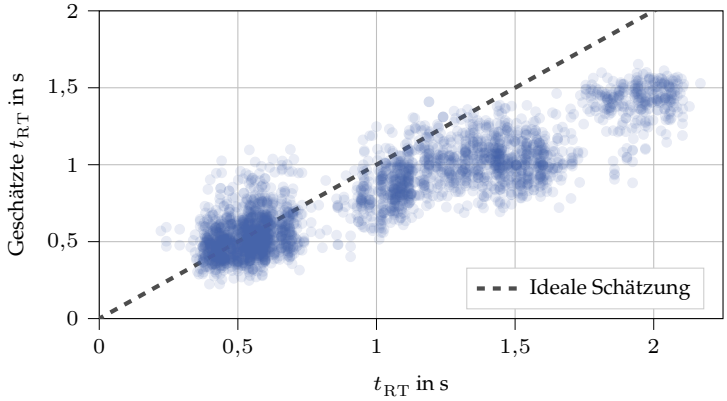


Abbildung 3.7 Verteilung der geschätzten Nachhallzeiten in Bezug auf die berechneten Referenzwerte der Testdaten aller neun Trainings- und Testräume des RIR-Datensatzes [135].

größeren t_{RT} das Zusammenspiel zwischen Raumklang und Instrumentencharakteristik des Klaviers eine bedeutendere Rolle, da beim Klavier ein Nachhalleffekt gewünscht ist und dieser nur schwer vollständig vom Raumklang zu trennen ist. Folglich scheint das KNN den Einfluss des Klaviers in seiner Schätzung implizit etwas höher zu bewerten. Trotz der systematischen Unterschätzung sind die Schätzwerte für große Nachhallzeiten im Mittel höher als für kleine, sodass eine Einschätzung des Nachhalls einer unbekannten Aufnahme möglich ist.

3.3 Anwendung zur Einordnung der Transkriptionsgüte

Die in Abschnitt 3.2 vorgestellten Qualitätsparameter können zur allgemeinen Bewertung der Aufnahmequalität eines Musiksignals, aber auch zur gezielten Einschätzung hinsichtlich der erreichbaren Güte einer MIR-Aufgabe bei störungsbehafteten Eingangssignalen verwendet werden. Dafür muss der Zusammenhang zwischen den charakteristischen Qualitätsparametern und den Ergebnissen der MIR-Aufgabe bekannt sein. Da die in dieser Arbeit definierten objektiven Qualitätsparameter

auf keine spezifische Aufgabe zugeschnitten sind, können sie theoretisch zur Einordnung der Güte zahlreicher MIR-Algorithmen eingesetzt werden. Eine speziell für Klaviersignale häufig erforschte Aufgabe ist die Extraktion der gespielten Noten einer Musikaufnahme, die sogenannte Transkription. Aus diesem Grund wird die entwickelte Qualitätsbewertung hier exemplarisch am Beispiel der Klaviertranskription angewendet, die über den in der Literatur etablierten Transkriptionsansatz *Onsets and Frames* [44] umgesetzt wird.

Wie in den vorherigen Abschnitten bilden die ersten 30 s der 60 realen Klavieraufnahmen des MAPS-Datensatzes die Basis zur Generierung der Testdaten. Sie werden, wie in den Abschnitten 3.2.1 bis 3.2.3 beschrieben, mit Störgeräuschen definierter Energie überlagert oder aufgenommenen Raumimpulsantworten gefaltet, um unterschiedliche Aufnahmebedingungen mit vorgegebenen Werten der Qualitätsparameter zu erzeugen.

Zur Bewertung von Transkriptionsergebnissen werden die extrahierten Einzelnoten in der Regel den Kategorien korrekt detektiert (*true positiv*, TP) oder falsch detektiert (*false positiv*, FP) zugeordnet und die nicht detektierten Noten des Musikstücks als *false negativ* (FN) erfasst. Daraus wird das für binäre Klassifikationsaufgaben etablierte F-Maß

$$F = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{3.2}$$

als Gütemaß der Transkription bestimmt. Eine Note wird dabei als korrekt erkannt klassifiziert, wenn der Zeitpunkt ihres Anschlags maximal 50 ms von dem bekannten exakten Anschlagszeitpunkt und die detektierte Notenfrequenz maximal 50 cent von der richtigen Frequenz der Note abweicht [107]. Die zusätzliche Berücksichtigung des Notenendes wird bei der Klaviertranskription häufig weggelassen [44], weil der genaue Zeitpunkt des Endes aufgrund der ausklingenden Charakteristik eines Klaviertons nur schwer zu detektieren ist. Deshalb wird auch in dieser Arbeit darauf verzichtet. Um die durch veränderte Aufnahmebedingungen hervorgerufenen Veränderungen in der Transkriptionsgüte hervorzuheben, wird in dieser Arbeit das relative F-Maß

$$F_{\mathrm{rel}} = \frac{F}{F_{\mathrm{orig}}} = \frac{\mathrm{TP} \cdot \left(\mathrm{TP}_{\mathrm{orig}} + 0.5\left(\mathrm{FP}_{\mathrm{orig}} + \mathrm{FN}_{\mathrm{orig}}\right)\right)}{\left(\mathrm{TP} + 0.5\left(\mathrm{FP} + \mathrm{FN}\right)\right) \cdot \mathrm{TP}_{\mathrm{orig}}} \tag{3.3}$$

anstelle des normalen F-Maßes angegeben. Es bezieht das normale F-Maß auf das F-Maß F_{orig} des nicht überlagerten Originalsignals. Daher bedeutet ein F_{rel} von 100 %, dass die Transkriptionsgüte im betrachteten Fall genauso hoch ist wie für das originale Musiksignal.

Im ersten Schritt wird der Einfluss von Hintergrundrauschen auf die Transkriptionsgüte analysiert. Dazu werden die SNR-Werte der mit unterschiedlichen Arten von Hintergrundrauschen überlagerten Klavieraufnahmen mithilfe des KNN aus Abschnitt 3.2.1 geschätzt. Darüber hinaus werden die überlagerten Klavieraufnahmen durch den Algorithmus *Onsets and Frames* transkribiert. Die Verteilungen der resultierenden Datenpunkte aus relativem F-Maß und SNR-Schätzwert sind in Abbildung 3.8 für jede Rauschart separat aufgeführt. Aus diesen Datenpunkten ergeben sich die Kurven der mittleren relativen F-Maße pro SNR-Stufe der Testdaten bezogen auf den Mittelwert der dafür geschätzten SNR-Werte, die in schwarz dargestellt sind. Da die SNR-Schätzung in Abschnitt 3.2.1 für die betrachteten Klavieraufnahmen nur kleine mittlere Fehler aufwies und damit erfolgreich validiert wurde, reicht die Untersuchung der SNR-Schätzwerte aus. Des Weiteren wurde die ähnliche Charakteristik der Kennlinien bezogen auf geschätzte und vorgegebene SNR-Werte in Vorversuchen bestätigt. Die Extreme des SNR-Wertebereichs werden insgesamt seltener geschätzt, sodass die schwarzen Kennlinien nicht ganz bis −5 dB oder 20 dB reichen und leicht in SNR-Richtung gestaucht sind.

Aus den Diagrammen in Abbildung 3.8 wird deutlich, dass die Transkriptionsergebnisse nur marginal von überlagertem braunen Rauschen beeinflusst werden. Erst unterhalb eines SNR von 3 dB reduziert sich das mittlere relative F-Maß im Vergleich zum nicht verrauschten Fall. Die Überlagerung mit Rauschen der drei anderen Rauschtypen führt dagegen zu einer deutlichen Verschlechterung der Transkription. Bei hochfrequentem Rauschen besteht im betrachteten Wertebereich zwischen dem geschätzten SNR-Wert und dem in der Transkription erzielten mittleren relativen F-Maß ein annähernd linear Zusammenhang. Im Falle von weißem und pinkem Rauschen bleiben die Transkriptionsergebnisse für ein SNR zwischen 10 dB und 20 dB näherungsweise konstant. Für SNR-Werte unter 10 dB verschlechtern sie sich aber deutlich, wobei ein fester Proportionalitätsfaktor approximiert werden kann. Daraus folgt, dass weißes, pinkes und hochfrequentes Hintergrundrauschen einen

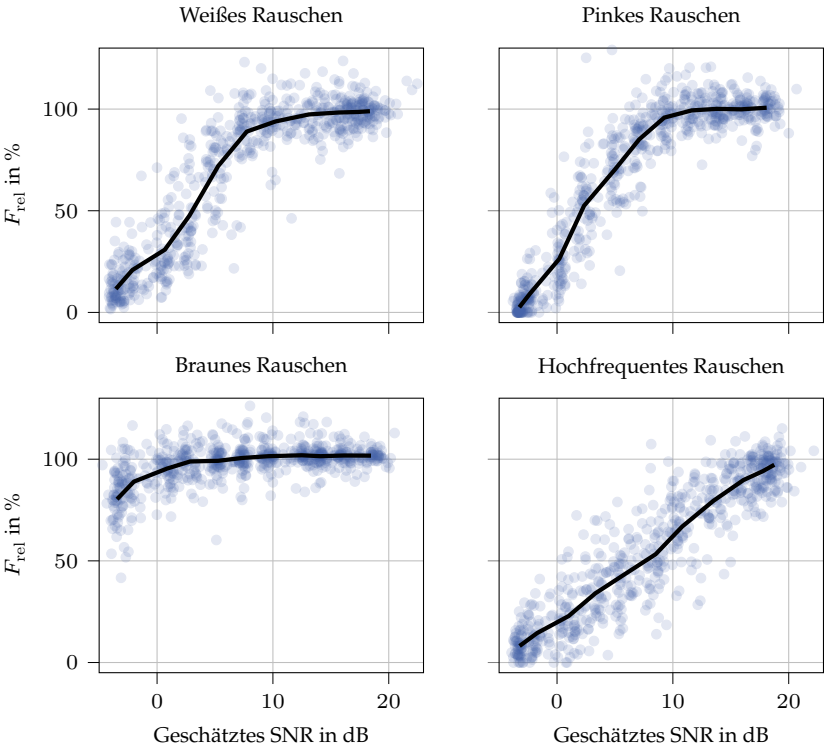


Abbildung 3.8 Verteilungen der Transkriptionsergebnisse in Abhängigkeit der geschätzten SNR-Werte bei überlagertem Hintergrundrauschen. Die schwarzen Linien repräsentieren die mittleren relativen F-Maße über den geschätzten SNR-Mittelwerten jeder SNR-Stufe.

großen Einfluss auf die Güte der Klaviertranskription haben und die entwickelte SNR-Schätzung deshalb zur Abschätzung der erzielbaren Transkriptionsergebnisse verwendet werden kann.

Einzelne Datenpunkte weichen von den beschriebenen Zusammenhängen stark ab. Diese Ausreißer treten sowohl nach oben als auch nach unten auf und können unter anderem durch die verschiedenen Klavieraufnahmen des MAPS-Datensatzes erklärt werden, die unterschiedlich schwer zu transkribieren sind. Darüber hinaus besitzen mehrere Datenpunkte relative F-Maße von über 100 %, was bedeutet, dass das

mit Hintergrundrauschen überlagerte Musiksignal bessere Transkriptionsergebnisse erzielt als die Originalaufnahme. Der Hauptgrund für diesen zunächst überraschenden Effekt ist die in der Regel reduzierte Anzahl an detektierten Noten bei Überlagerung von Rauschen. Werden in der originalen Musikaufnahme zu viele Noten erkannt, kann sich das Transkriptionsergebnis durch eine geringere Anzahl detektierter Noten verbessern, sofern durch das überlagerte Rauschen mehrheitlich falsch erkannte Noten nicht mehr detektiert werden.

Zur Analyse der Auswirkungen von zeitvarianten Störgeräuschen auf die Klaviertranskription werden die 30 s langen Klavieraufnahmen der MAPS-Teststücke analog zur Darstellung in Abbildung 3.4 mit drei jeweils 4 s langen Störsignalen einer Geräuschklasse überlagert. Folglich ist das zu transkribierende Musiksignal nur 40 % seiner Dauer mit kurzen Störgeräuschen überlagert, die Transkription wird aber trotzdem über die gesamten 30 s durchgeführt. In Abbildung 3.9 sind alle aus der Klaviertranskription dieser Testdaten resultierenden Verteilungen des relativen F-Maßes in Abhängigkeit der SNR-Schätzwerte kurzer Abschnitte dargestellt. Das geschätzte SNR entspricht dabei dem Mittelwert der SNR-Schätzungen aller 2-sekündigen Zeitabschnitte eines Musiksignals, in dem das jeweils angegebene Störgeräusch mindestens zeitweise enthalten ist. In schwarz sind die Mittelwerte des relativen F-Maßes pro SNR-Stufe der generierten Testdaten illustriert, welche über die jeweiligen Mittelwerte der SNR-Schätzungen aufgetragen sind. Wie schon hinsichtlich Abbildung 3.4 diskutiert, werden hohe SNR-Werte im Mittel deutlich unterschätzt und niedrige SNR-Werte überschätzt. Deshalb reichen die schwarzen Kennlinien in Abbildung 3.9 nicht bis –5 dB oder 20 dB, sondern sind sogar noch weiter in SNR-Richtung gestaucht als die Kennlinien für Hintergrundrauschen in Abbildung 3.8.

Die Zusammenhänge zwischen den SNR-Schätzwerten der zeitvarianten Störgeräusche und den dadurch hervorgerufenen Verschlechterungen der Transkriptionsgüte sind für alle untersuchten Geräuschklassen vergleichbar. Im Mittel können sie durch lineare Kennlinien angenähert werden, wobei das relative F-Maß bei niedrigen SNR-Werten deutlich weniger abfällt als im Falle des Hintergrundrauschens. Das liegt vor allem daran, dass nur 40 % der Signallänge von Störgeräuschen überlagert ist, sodass die Transkriptionsgüte der ungestörten Passagen nicht beeinflusst

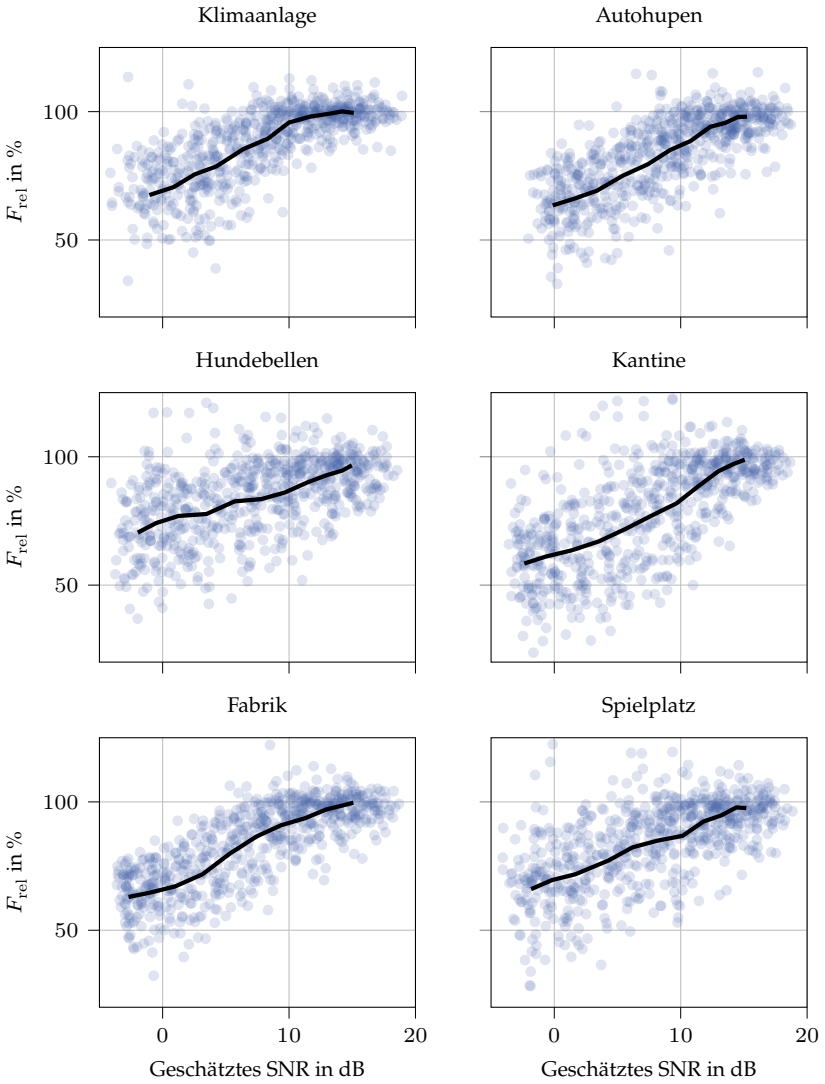


Abbildung 3.9 Verteilungen der Transkriptionsergebnisse in Abhängigkeit der geschätzten SNR-Werte bei überlagerten zeitvarianten Störgeräuschen verschiedener Klassen (Störung über 40 % der Signallänge, wie in Abbildung 3.4). Die schwarzen Linien repräsentieren die mittleren relativen F-Maße über den geschätzten SNR-Mittelwerten jeder SNR-Stufe.

Tabelle 3.5 Geschätzte SNR-Werte für (Hintergrundrauschen; kurze Störgeräusche) in dB bei gleichzeitiger Überlagerung von zufälligem Hintergrundrauschen und zeitvarianten Störgeräuschen aller betrachteten Rauscharten mit definierten SNR-Stufen.

		SNR der zeitvarianten Störgeräusche		
		15 dB	7,5 dB	0 dB
SNR des	15 dB	(19,5; 11,8)	(18,5; 7,5)	(17,1; 1,4)
Hintergrund-	7,5 dB	(11,3; 7,9)	(10,9; 5,3)	(11,5; 0,5)
rauschens	0 dB	(2,9; 1,9)	(2,6; 0,9)	(2,7; –1,5)

wird. Abhängig vom Musiksignal während der Störgeräusche wird das zugehörige Transkriptionsergebnis unterschiedlich stark beeinträchtigt, da Notenanfänge am wichtigsten und ausklingende Noten nahezu irrelevant für eine erfolgreiche Transkription sind. Darüber hinaus gelten auch in diesem Fall die bereits diskutierten Gründe für Ausreißer und relative F-Maße über 100 %. Aus diesen Gründen weisen die ermittelten Datenpunkte für zeitvariante Störgeräusche eine größere Varianz als die bei Hintergrundrauschen auf. Trotzdem kann ein klarer Zusammenhang zwischen den geschätzten SNR-Werten und der resultierenden Transkriptionsgüte hergestellt werden. Folglich lässt sich die erzielbare Güte der Klaviertranskription anhand einer Bewertung der Aufnahmequalität durch die SNR-Schätzung abschätzen.

Im Gegensatz zu den bisher untersuchten Störungen durch eine Geräuschart können in realen Aufnahmen gleichzeitig mehrere Arten auftreten. Ist Hintergrundrauschen vorhanden, kann es z. B. zu Überlagerungen mit kurzen Störgeräuschen kommen. Daher werden die Schätzergebnisse der KNN für Hintergrundrauschen und kurze Störgeräusche bei ausgewählten Kombinationen von SNR-Stufen der beiden Störklassen analysiert. Die zugehörigen Ergebnisse sind in Tabelle 3.5 aufgelistet. Aus Gründen der Übersichtlichkeit werden nur drei der elf untersuchten SNR-Stufen angegeben, wobei die Schätzwerte für kurze Störgeräusche die SNR-Schätzungen aller durch zeitvariante Geräusche überlagerten Zeitabschnitte der Länge 2 s repräsentieren. Die Schätzung des SNR bei Hintergrundrauschen wird so gut wie gar nicht von der Präsenz kurzer Störgeräusche beeinflusst, da die mittleren Schätzwerte auf jeder

betrachteten SNR-Stufe des Hintergrundrauschens nahezu konstant bleiben. Auch wenn die Schätzungen für alle Stufen etwas zu hoch sind, können die Abstufungen gut unterschieden werden. Folglich wird in der SNR-Schätzung ausschließlich Rauschen mit stationärer Charakteristik berücksichtigt, womit das Ziel des auf Hintergrundrauschen ausgelegten KNN erfüllt ist. Dagegen wird die Schätzung des SNR bei kurzen Störgeräuschen stark durch überlagertes Hintergrundrauschen beeinträchtigt, weil das KNN maximal die SNR-Werte des Hintergrundrauschens schätzt. Da diese Schätzung aber auch alle Störgeräusche innerhalb eines kurzen Zeitabschnitts berücksichtigen soll, entspricht dieses Verhalten der Vorgabe. Die Präsenz von zusätzlichen kurzen Störgeräuschen mit hoher Signalenergie schlägt sich in den Ergebnissen durch noch niedrigere SNR-Werte nieder. Diese Ergebnisse bestätigen die Möglichkeit einer Bewertung der Aufnahmequalität eines Musiksignals mit überlagertem Hintergrundrauschen und kurzen Störgeräuschen durch die beiden entwickelten, separaten KNN zur SNR-Schätzung.

Als weitere Komponente der Qualitätsbewertung von Musikaufnahmen wird der Einfluss der Raumakustik auf die Klaviertranskription anhand der Anfangsnachhallzeit t_{EDT} untersucht. Aufgrund der genaueren Schätzergebnisse in Abschnitt 3.2.3 wird sie gegenüber der Nachhallzeit t_{RT} bevorzugt. In Abbildung 3.10 ist die resultierende Verteilung der relativen F-Maße über die geschätzten Anfangsnachhallzeiten (bezogen auf einen Energieabfall um 60 dB) dargestellt. Insgesamt kann dem Diagramm anhand der Mittelwertkennlinie ein Zusammenhang zwischen langen Anfangsnachhallzeiten und schlechten Transkriptionsergebnissen entnommen werden. Die Varianz der Daten ist aber sehr hoch, sodass eine Vorhersage der Transkriptionsgüte auf Basis einzelner Schätzungen der t_{EDT} nur sehr eingeschränkt möglich ist. Wie schon in Abschnitt 3.2.3 erläutert, spielt die Charakteristik des Klaviers, welches in der Klangentstehung einen gewünschten Nachhalleffekt hervorruft, dabei eine große Rolle. Durch diesen immer präsenten Nachhall ist der Einfluss von Raumakustik mit kleiner Nachhallzeit im resultierenden Klaviersignal nicht signifikant und erschwert die Parameterschätzung. Daher ist für die Bewertung der Transkriptionsgüte bei Klaviersignalen nur eine grobe Klassifikation der Anfangsnachhallzeiten in kurz und lang sinnvoll. Mithilfe der Punktwolke in Abbildung 3.10 wird ein

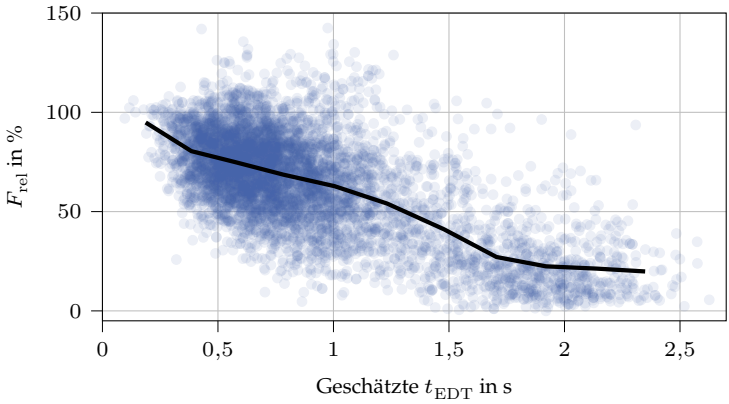


Abbildung 3.10 Verteilungen der Transkriptionsergebnisse in Abhängigkeit der geschätzten Anfangsnachhallzeiten bei Faltung mit RIRs aller neun Räume des RIR-Datensatzes [135]. Die schwarze Kennlinie repräsentiert die mittleren relativen F-Maße über dem gleitenden Mittelwert der geschätzten t_{EDT}.

Schwellenwert bei 1,2 s festgelegt, da unter dieser Schwelle eine große Datenwolke mit großer Varianz liegt und das mittlere F-Maß an dieser Grenze ungefähr den Mittelwert seines Wertebereichs annimmt.

Abschließend wird der Zusammenhang zwischen den Ergebnissen der Klaviertranskription und den gekoppelten Beeinträchtigungen aus Hintergrundrauschen, kurzen Störgeräuschen und unterschiedlichen Raumimpulsantworten analysiert. Die resultierenden relativen F-Maße sind in Tabelle 3.6 in Abhängigkeit der Schätzungen der Anfangsnachhallzeiten sowie den schon in Tabelle 3.5 betrachteten SNR-Stufen von Hintergrund- und kurzen Störgeräuschen aufgeführt. Dabei werden die Schätzwerte der t_{EDT} nicht exakt betrachtet, sondern wie oben beschrieben nur eine Klassifikation in zwei Klassen durchgeführt. Durch die Transkriptionsergebnisse wird der für einzelne Störungen analysierte Zusammenhang zwischen geschätzten SNR-Werten und Transkriptionsgüte auch für den Fall von gekoppelten Beeinträchtigungen bestätigt, da die Ergebnisse im Falle von zeitvarianten kurzen Störgeräuschen etwas schlechter werden, für stärkeres Hintergrundrauschen aber stark abfallen. Darüber hinaus wird der negative Einfluss von großen Anfangsnachhallzeiten anhand der in jedem Fall schlechteren F-Maße für die

Tabelle 3.6 Relatives F-Maß der Klaviertranskription in Abhängigkeit des geschätzten t_{EDT} und verschiedener SNR-Stufen von Hintergrundrauschen und zeitvarianten Störgeräuschen. Die Testsignale beinhalten gleichzeitig alle drei definierten Störeinflüsse.

	SNR des Hintergrundrauschens	SNR der zeitvar. Störgeräusche 15 dB	7,5 dB	0 dB
$t_{\mathrm{EDT}} < 1{,}2\,\mathrm{s}$	15 dB	65 %	58 %	49 %
	7,5 dB	46 %	43 %	40 %
	0 dB	9 %	9 %	12 %
$t_{\mathrm{EDT}} \geq 1{,}2\,\mathrm{s}$	15 dB	34 %	31 %	25 %
	7,5 dB	19 %	18 %	15 %
	0 dB	4 %	4 %	4 %

Klasse $t_{\mathrm{EDT}} \geq 1{,}2\,\mathrm{s}$ deutlich. Daher ist die vorgeschlagene Klassifikation anhand der t_{EDT} mit der Schwelle 1,2 s sinnvoll.

Aufgrund der durch mehrere Einflussfaktoren reduzierten Aufnahmequalität sind die in Tabelle 3.6 angegebenen Ergebnisse der Klaviertranskription schlechter als nur bei einer Art der Beeinflussung. Nach Abbildung 3.10 trägt dazu vor allem das im Mittel deutlich reduzierte relative F-Maß für den Fall einer veränderten Raumakustik bzw. unterschiedlicher Anfangsnachhallzeiten bei. Nichtsdestotrotz ist ein eindeutiger Zusammenhang zwischen der Transkriptionsgüte und den SNR-Schätzungen sowie der Klassifikation anhand der geschätzten t_{EDT} festzustellen. Deshalb kann die aus einer geringen Aufnahmequalität resultierende niedrige Transkriptionsgüte mithilfe der vorgestellten Qualitätsparameter für Hintergrundrauschen, kurze Störgeräusche und Raumimpulsantwort abgeschätzt und vorhergesagt werden. Des Weiteren ist anhand der mit unabhängigen KNN geschätzten Parameter eine Einschätzung potenzieller Gründe für die reduzierte Güte möglich.

4 Zeitabhängige Instrumentendetektion

Die in einer Musikaufnahme zu hörenden Musikinstrumente sind in der Regel nicht vorab bekannt. Sie stellen aber eine wichtige Eigenschaft zur Beschreibung und Einordnung von Musiksignalen dar. Gerade bei polyphonen Instrumentalstücken, z. B. von Ensembles unterschiedlicher Instrumentengattungen, dienen sie als ein charakteristisches Merkmal. Darüber hinaus erleichtert die Information über aktive Instrumente viele MIR-Aufgaben wie die Erkennung des Genres oder der Gemütslage eines Musikstücks, die Transkription der vorkommenden Noten oder die Separation einzelner Instrumentenstimmen.

In der Literatur werden diverse Ansätze zur Detektion der in einem Musiksignal präsenten Musikinstrumente vorgeschlagen, teilweise zur Merkmalsextraktion, aber auch als Vorverarbeitung für andere MIR-Aufgaben. Abschnitt 4.1 gibt einen Überblick über die relevantesten Verfahren. Im Gegensatz zu vielen Literaturansätzen erfolgt die in dieser Arbeit vorgeschlagene Erkennung aktiver Instrumente zeitabhängig. Dabei werden neben den Beträgen der Zeit-Frequenz-Darstellungen auch Phaseninformationen in den Eingangsdaten genutzt. Diese werden in Form von speziellen Zeit-Frequenz-Darstellungen eingespeist, die in Abschnitt 4.2 vorgestellt werden. Die Schätzung der aktiven Instrumente erfolgt mithilfe von unterschiedlichen KNN, deren Architekturen in Abschnitt 4.3 beschrieben sind. Ihre Auswirkungen auf die Güte der Instrumentendetektion sowie der Einfluss von zusätzlich eingespeisten Phaseninformationen werden in Abschnitt 4.4 untersucht.

Das vorliegende Kapitel beinhaltet bereits veröffentlichte Erkenntnisse zum Einfluss von Phasendarstellungen [A2] und unterschiedlichen Zeit-Frequenz-Darstellungen [A3] auf die zeitabhängige Instrumentendetektion. Darin wird die in Abschnitt 4.3 vorgestellte Netzarchitektur mit zweidimensionaler Einspeisung analysiert.

4.1 Stand der Forschung

Der Klang eines Musikinstruments hängt, wie in Abschnitt 2.1.2 beschrieben, sowohl vom Intensitätsverhältnis der Grund- und Oberschwingungen als auch von variierenden Faktoren wie der Spieltechnik ab. Durch diese hohe Variabilität ist die automatische Detektion des in einem Musiksignal enthaltenen Instrumententyps eine komplexe MIR-Aufgabe, zu deren Lösung sich Verfahren des maschinellen Lernens anbieten und ein deterministischer Algorithmus meist nicht ausreicht. Schon im Jahre 1995 wurden ein KNN sowie ein Nächste-Nachbarn-Klassifikator vorgeschlagen, die eine Instrumentenerkennung für monophone Instrumentensignale innerhalb einer definierten Oktave realisieren [58]. Aus den unter Laborbedingungen aufgenommenen Signalen von vier Instrumenten wurden vor der Klassifikation Merkmale zur Signalenergie extrahiert und eine Hauptkomponentenanalyse durchgeführt. Auch die Instrumentenerkennung mit einem Gauß'schen Mischmodell und einer Support-Vektor-Maschine wird durch eine zusätzliche Hauptkomponentenanalyse der aus Soloaufnahmen extrahierten Merkmale verbessert [30]. Dabei umfassen die Soloaufnahmen sowohl Einzelnoten als auch längere Passagen von Holzblasinstrumenten. Eine stabilere Instrumentenerkennung bei isolierten Einzelnoten kann durch die Verwendung von *Hidden-Markov*-Modellen mit Interpolation [148] erzielt werden.

Bei der Instrumentendetektion von mehrstimmigen Musikaufnahmen spielen Überlappungen in den Oberschwingungen gleichzeitig gespielter Instrumententöne eine große Rolle. Ihr Einfluss auf die Detektion kann mithilfe einer Gewichtung der aus dem Signal extrahierten Merkmale reduziert werden [68]. Deutlich bessere Detektionsergebnisse werden gerade für polyphone Musiksignale durch die Anwendung von KNN wie dem tiefen CNN von Han et al. [42] erzielt. Dieses Netz zur Erkennung des in einem Musikausschnitt dominanten Instruments verwendet an seinem Eingang das Mel-Spektrogramm des Musiksignals, welches, wie in Abschnitt 2.2.1 beschrieben, eine STFT mit angepasster Frequenzskala darstellt. Mehrere Schätzungen kurzer, aufeinanderfolgender Zeitabschnitte werden zu einer Instrumentenschätzung für das Gesamtsignal zusammengefasst. Um solche CNN möglichst effizient auszulegen, haben Pons et al. [104] eine Entwurfsstrategie für CNN-Achitekturen mit Einspeisung von Spektrogrammen entwickelt, die relevante Merkmale

hinsichtlich der Klangfarbe berücksichtigt, aber nur mit wenigen, musikalisch motivierten Filterkernen auskommt.

Spektrogramme enthalten nur Betragsinformationen der ursprünglichen Zeit-Frequenz-Darstellungen, weshalb die zugehörigen Phaseninformationen nicht berücksichtigt werden. Um diese ebenfalls in die Instrumentendetektion einzubeziehen, können sie in Form der modifizierten Gruppenlaufzeit (ModGD), die eine Art Phasenspektrogramm repräsentiert und aus der STFT berechnet wird, eingespeist werden [24]. Der zugehörige Ansatz zur Instrumentenerkennung basiert dabei auf einem Gauß'schen Mischmodell, dessen Detektionsergebnisse sich durch die Hinzunahme der ModGD verbessern. Zusätzliche Phasendarstellungen werden auch zur verbesserten Quellentrennung von Musiksignalen erfolgreich eingesetzt [121]. In diesem Fall werden die Daten der ModGD aber noch weiter verarbeitet, um eine an das entwickelte RNN angepasste Phasendarstellung zu generieren.

Neben der Einspeisung von Zeit-Frequenz-Darstellungen des betrachteten Musiksignals, die ein gewisses Vorwissen über die Zusammensetzung von Musiksignalen beinhalten, ist auch eine direkte Einspeisung des Zeitsignals möglich. Dieleman und Schrauwen [21] haben diese Variante der Ende-zu-Ende-Architektur eines CNN-Modells vorgestellt. Neben der Erkennung der im gesamten Musiksignal aktiven Instrumente detektiert der Ansatz weitere charakteristische Eigenschaften wie Genre oder Besetzung der Aufnahme. Im Vergleich zur Einspeisung von Spektrogrammen muss beim Ende-zu-Ende-Ansatz keine Vorverarbeitung durchgeführt werden, aber die resultierenden Detektionsergebnisse sind im vorgestellten Fall etwas schlechter. Darüber hinaus werden in der ersten CNN-Schicht Filter gelernt, deren Werte Zeitsignalen mit einer dominanten Sinusschwingung entsprechen. Folglich ähnelt die Operation der ersten CNN-Schicht der Berechnung von Spektrogrammen. Diese Ergebnisse bestätigen sich auch für eine reine Instrumentenerkennung mit CNN-Modell und direkter Signaleinspeisung [76].

Wie bei allen lernenden Verfahren beeinflusst die im Training vorhandene Datenmenge und -diversität das erreichbare Ergebnis. Durch Augmentierungsstrategien wie das Mischen von synchronisierten Audiosegmenten hinsichtlich Tonhöhe, Tempo oder Genre lassen sich bestehende Datensätze aus monophonen Instrumentenaufnahmen zu po-

lyphonen Datensätzen erweitern [70]. Dabei wird der Mischvorgang mithilfe von CQT-Darstellungen der Einzelsequenzen durchgeführt. Alternativ kann bei kleinen Mengen gelabelter Daten halbüberwachtes Lernen angewendet werden [23]. Ein weiteres Problem in vorhandenen Trainingsdatensätzen stellen schwach gelabelte Daten dar, in denen aktive Instrumente immer wieder als nicht aktiv erscheinen. Abhilfe dagegen schaffen *Attention*-Mechanismen im Detektionsmodell, die aus den Musiksignalen zusätzliche Informationen über die zeitliche Aktivität von charakteristischen Merkmalen extrahieren [40]. Daraus lassen sich die für die Musikinstrumente relevanten Zeitabschnitte innerhalb der betrachteten 10-sekündigen Aufnahme identifizieren und folglich die Detektion der Instrumente verbessern.

Die Ähnlichkeit zwischen verwandten Instrumenten wird normalerweise nur implizit in den gelernten Modellparametern berücksichtigt, kann aber auch durch eine feste Taxonomie und die daraus resultierende hierarchische Klassifikation [31] vorgegeben werden. Mithilfe eines sogenannten *Few-Shot*-Lernverfahrens, das Zusammenhänge auch schon aus wenigen Lernbeispielen identifiziert, lassen sich die Merkmale der vorgegebenen Hierarchiestruktur automatisch lernen [35]. Der Vorteil dabei ist die flexible Erweiterbarkeit der Struktur und die gute Anwendbarkeit auf Instrumente mit wenigen Trainingsbeispielen oder auf im Training unbekannte Instrumente, die aufgrund der Hierarchie besser bzw. zumindest grob klassifiziert werden können.

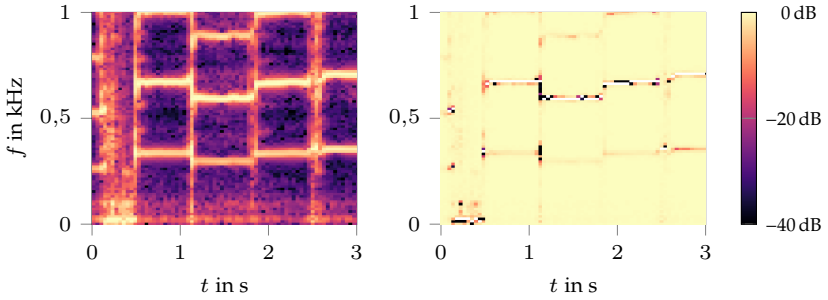
Alle bisher in diesem Kapitel vorgestellten Verfahren zur Instrumentendetektion schätzen die in der betrachteten Musikaufnahme zu hörenden Instrumente zeitunabhängig. Da die in die Modelle eingespeisten Signalabschnitte eine Länge von mindestens 1 s besitzen, kann die Instrumentenaktivität nur mit einer Zeitabhängigkeit in dieser Größenordnung angegeben werden. Feinere Auflösungen der Instrumentendetektion sind durch zeitabhängige Schätzungen der aktiven Instrumente möglich, die nach dem Training mit zeitunabhängigen Daten über eine Modifikation der letzten CNN-Schicht erreicht werden können [79]. Sind auch ausreichend zeitabhängige Trainingsdaten vorhanden, liefern damit trainierte CNN-Modelle bessere Ergebnisse. Ein solches Modell stellt der Ansatz nach Hung und Yang [53] dar, wobei das CNN in diesem Fall mit der Betragsdarstellung einer CQT gespeist wird. Zusätzlich erhält das Modell

die in einem separaten Ansatz detektierten Tonhöheninformationen des betrachteten Musiksignals. Um diese beiden Schritte nicht nacheinander durchführen zu müssen, kann ein gemeinsames Modell zur kombinierten zeitabhängigen Schätzung der aktiven Musikinstrumente und Töne eingesetzt werden [51]. Das Modell profitiert dabei von der im Training erlernten Wechselwirkung zwischen Instrumentenklang und Tonhöhe, indem es diese für die Schätzung beider Teilaufgaben nutzt.

4.2 Zeit-Frequenz-Darstellungen von Phaseninformation

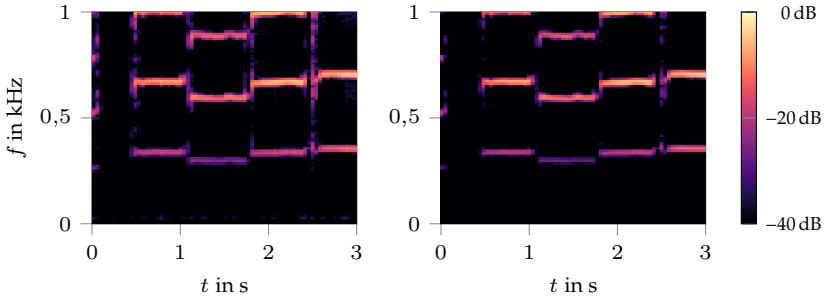
Viele MIR-Algorithmen, wie beispielsweise auch die in Kapitel 3 vorgestellte Qualitätsbewertung, nutzen nur den Absolutbetrag $|X[m,k]|$ der aus dem jeweiligen Musiksignal berechneten Zeit-Frequenz-Darstellung. In den Literaturansätzen zur Instrumentenerkennung wird sehr häufig der Betrag der STFT oder eine daraus abgeleitete Darstellung wie das Mel-Spektrogramm eingesetzt, deren Frequenzauflösung durch die konstante Schrittweite zwischen den Frequenzbins der STFT begrenzt ist. Deshalb und aufgrund der nach Gleichung (2.9) exponentiell ansteigenden Notenfrequenzen können kleine Tonunterschiede in tiefen Frequenzlagen nicht mehr in den Spektrogrammen detektiert werden. Die exakte Information der damit verbundenen Frequenzabweichung ist in der Phase $\theta_{\mathrm{STFT}}[m,k]$ der STFT enthalten. Wegen ihrer Periodizität mit 2π ist die Phase aber nicht direkt interpretierbar und muss erst entfaltet werden, da sonst Mehrdeutigkeiten auftreten können.

Im Folgenden werden drei Ansätze für Zeit-Frequenz-Darstellungen der Phaseninformation beschrieben, welche die zusätzlichen Informationen in interpretierbare Matrizen überführen. Diese Matrizen besitzen die gleichen Dimensionen wie die STFT, wodurch sie parallel zum Spektrogramm in das Modell zur Instrumentendetektion eingespeist werden können. Die resultierenden Darstellungen aller drei Phasenansätze sowie ihrer Ausgangsdarstellungen, der STFT und der in Abschnitt 4.2.1 verwendeten Gruppenlaufzeit, sind in Abbildung 4.1 für den Ausschnitt eines Trompetensignals veranschaulicht.



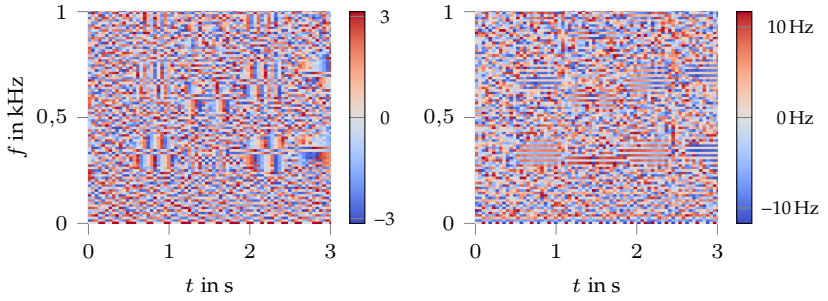
(a) Spektrogramm der STFT.

(b) Gruppenlaufzeiten.



(c) Modifizierte Gruppenlaufzeiten.

(d) Produktspektrum.



(e) Phasenwerte der STFT.

(f) Frequenzfehlermatrix.

Abbildung 4.1 Zeit-Frequenz-Darstellungen aus Betrags- und Phaseninformationen der STFT des identischen Ausschnitts einer Trompetenaufnahme (Musikstück 18_Nocturne) aus dem URMP-Datensatz [75].

4.2.1 Modifizierte Gruppenlaufzeit

Ein Ansatz zur Verarbeitung von Phaseninformation ist die Betrachtung der Gruppenlaufzeit (englisch *group delay*, GD). Sie wird häufig zur Bestimmung der Verzögerung von Wellenpaketen eingesetzt und entfaltet die Phasenwerte implizit, sodass keine Phasensprünge mehr auftreten. Im Kontinuierlichen ist die Gruppenlaufzeit über

$$\tau_{\mathrm{GD}}(\omega) = -\frac{\mathrm{d}}{\mathrm{d}\omega}\theta(\omega) = -\operatorname{Im}\left\{\frac{\mathrm{d}}{\mathrm{d}\omega}\log\left(X_{\mathrm{FT}}(\omega)\right)\right\} \tag{4.1}$$

definiert [2], wobei X_{FT} die Fourier-Transformierte des betrachteten Signals x und ω die Kreisfrequenz repräsentieren. Da eine STFT im Diskreten aus aneinandergereihten Fourier-Transformationen besteht, kann die Berechnung auf den Fall einer diskreten, zeit- und frequenzabhängigen Gruppenlaufzeit erweitert werden. Daraus lässt sich die Formel

$$\tau_{\mathrm{GD}}[m,k] = \frac{X_{\mathrm{Re}}[m,k]\;\Xi_{\mathrm{Re}}[m,k] + X_{\mathrm{Im}}[m,k]\;\Xi_{\mathrm{Im}}[m,k]}{|X[m,k]|^2} \tag{4.2}$$

herleiten [24], in der X und Ξ die STFT-Transformierten der Signale $x[n]$ und $\xi = n \cdot x[n]$ darstellen. Die komplexen Werte der Transformierten werden aufgetrennt in ihre Real- (Re) und Imaginärteile (Im) verwendet. Aufgrund der besseren Lesbarkeit wird in diesem Abschnitt auf den Index ‚STFT' aller Transformierten verzichtet.

Um $\tau_{\mathrm{GD}}[m,k]$ berechnen zu können, darf der Betrag $|X_{\mathrm{STFT}}[m,k]|$ laut Gleichung (4.2) nicht Null sein. Aber auch schon im Bereich nahe der Nullstellen von $X_{\mathrm{STFT}}[m,k]$ treten Probleme auf, da die kleinen Werte für $|X_{\mathrm{STFT}}[m,k]|$ im Nenner sehr hohe Gruppenlaufzeiten verursachen. Dadurch wird die eigentlich relevante Phaseninformation überlagert und die damit zusammenhängende Resonanzstruktur der Gruppenlaufzeit in diesem Bereich ausgeblendet [24]. Solche ungewollten Nullstellen werden unter anderem durch die Fensterung in der Kurzzeit-Analyse hervorgerufen [121], wie sie auch in der Berechnung der STFT enthalten ist. Deshalb wird die STFT-Transformierte X_{STFT} im Nenner von Gleichung (4.2) in der Berechnung der modifizierten Gruppenlaufzeit (ModGD) durch das cepstral geglättete Spektrum X_{CS} ersetzt. Dieses spezielle Spektrum wird über die diskrete Cosinus-Transformation (DCT) berechnet, welche auf das logarithmierte Leistungsspektrum $|X_{\mathrm{STFT}}[m,k]|^2$ angewendet

wird. Zur Unterdrückung der Nullstellen werden nur der Gleichanteil sowie die ersten N_{CS} der resultierenden DCT-Koeffizienten behalten und daraus anschließend die inverse DCT berechnet [150]. Das Ergebnis ist das cepstral geglättete Leistungsspektrum $|X_{\mathrm{CS}}[m,k]|^2$, welches Nullstellen der STFT-Transformierten unterdrückt und in der Sprachsignalverarbeitung weit verbreitet ist.

Neben dem cepstral geglätteten Spektrum X_{CS} enthält die Definition der modifizierten Gruppenlaufzeit zwei Designparameter ρ und γ, mithilfe derer ihr Dynamikbereich kontrolliert werden kann. Die Berechnung der zeit- und frequenzabhängigen modifizierten Gruppenlaufzeit erfolgt damit über [24]

$$\tau_{\mathrm{ModGD}}[m,k] = \mathrm{sign}\left(\tilde{\tau}\left[m,k\right]\right) \cdot \left(\left|\tilde{\tau}\left[m,k\right]\right|\right)^{\rho} \tag{4.3}$$

mit der Hilfsgröße

$$\tilde{\tau}[m,k] = \frac{X_{\mathrm{Re}}[m,k]\ \Xi_{\mathrm{Re}}[m,k] + X_{\mathrm{Im}}[m,k]\ \Xi_{\mathrm{Im}}[m,k]}{|X_{\mathrm{CS}}[m,k]|^{2\gamma}}\,. \tag{4.4}$$

Analog zur Berechnung der Gruppenlaufzeit in Gleichung (4.2) stellen die Größen im Zähler von Gleichung (4.4) den Real- bzw. Imaginärteil der jeweiligen STFT-Transformierten dar.

In dieser Arbeit werden zur Berechnung des cepstral geglätteten Spektrums die ersten $N_{\mathrm{CS}} = 30$ Koeffizienten der DCT verwendet, da sich diese Zahl in der Literatur bewährt hat [150]. Für die Wahl der Dynamikparameter ρ und γ wurden in Vorversuchen Werte zwischen 0,1 und 1 anhand der Eigenschaften hinsichtlich Dynamik und Rauschen der resultierenden modifizierten Gruppenlaufzeiten untersucht. Aus der Analyse aller Musikstücke des verwendeten Datensatzes gingen die empirisch besten Werte $\rho = 0{,}4$ und $\gamma = 0{,}99$ hervor, welche in der gesamten Arbeit genutzt werden. Darüber hinaus werden alle hier eingesetzten Zeit-Frequenz-Darstellungen der modifizierten Gruppenlaufzeit auf ihr Maximum normiert und in den logarithmischen Wertebereich überführt, um die Varianz zwischen den Darstellungen der unterschiedlichen Musikstücke zu reduzieren.

Die Auswirkungen der vorgestellten Modifikationen auf die Zeit-Frequenz-Darstellung der Gruppenlaufzeit werden im Vergleich der Abbildungen 4.1(b) und 4.1(c) deutlich. In der modifizierten Gruppenlaufzeit

treten die Spektren der gespielten Töne sehr deutlich als Linien hoher Werte hervor, wohingegen diese Struktur in der normalen Gruppenlaufzeit nur zu erahnen ist. Des Weiteren besitzen die Bereiche der gespielten Töne in Abbildung 4.1(b) teilweise kleinere Werte als in ihrer Umgebung und sind an mehreren Stellen von einzelnen Punkten mit sehr niedrigen Werten überlagert, wodurch die oben beschriebene Ausblendung der Resonanzstruktur sichtbar wird. Im Gegensatz zum Spektrogramm in Abbildung 4.1(a) ist bei der modifizierten Gruppenlaufzeit deutlich weniger Rauschen zu erkennen, sodass die gespielten Trompetentöne wesentlich klarer erscheinen. Dies kann die Erkennung der charakteristischen Instrumentenspektren vereinfachen.

4.2.2 Produktspektrum

Neben der Modifizierung zur ModGD kann die in Gleichung (4.2) definierte zeit- und frequenzabhängige Gruppenlaufzeit auch mit anderen Merkmalen kombiniert werden, um aussagekräftige Zeit-Frequenz-Darstellungen zu erhalten. Eine Möglichkeit bietet das Produktspektrum (PS)

$$P[m,k] = |X_{\mathrm{STFT}}[m,k]|^2 \cdot \tau_{\mathrm{GD}}[m,k] \tag{4.5}$$

$$= X_{\mathrm{Re}}[m,k]\ \Xi_{\mathrm{Re}}[m,k] + X_{\mathrm{Im}}[m,k]\ \Xi_{\mathrm{Im}}[m,k] \tag{4.6}$$

nach Zhu und Paliwal [153], das die Gruppenlaufzeit mit dem Leistungsspektrum der STFT durch eine Multiplikation verknüpft. Das Produktspektrum enthält folglich sowohl Betrags- als auch Phaseninformationen. Die Berechnungsformel (4.6) mit den STFT-Transformierten X und Ξ der Signale $x[n]$ bzw. $\xi = n \cdot x[n]$ resultiert aus Gleichung (4.2). Wie im vorherigen Abschnitt werden die Indizes ‚STFT' bei den Real- und Imaginärteilen von X und Ξ aus Gründen der besseren Lesbarkeit weggelassen. Nach der Berechnung des Produktspektrums erfolgt in dieser Arbeit stets eine Normierung auf sein Maximum, da die absolute Lautstärke von Musikstücken für die Instrumentendetektion von untergeordneter Bedeutung ist. Anschließend werden die normierten Werte in eine logarithmische Darstellung konvertiert.

Im Vergleich zum Spektrogramm und der Gruppenlaufzeit treten die Grund- und Oberschwingungen der Trompetentöne im zugehörigen Produktspektrum (Abbildung 4.1(d)) deutlich stärker hervor, da es außer-

halb dieser Frequenzen nur sehr geringe Werte besitzt. Das Produktspektrum ähnelt der Darstellung der modifizierten Gruppenlaufzeit, wobei es noch weniger Artefakte an den Tonanfängen und Tonenden aufweist. Dadurch erscheint es noch klarer, was bei der Erkennung charakteristischer Instrumentenspektren hilfreich ist. Darüber hinaus werden Geräusche in Zeiträumen ohne aktive Töne, wie beispielsweise in der ersten halben Sekunde, fast vollständig unterdrückt.

4.2.3 Frequenzfehlermatrix

Aufgrund der Zeit-Frequenz-Unschärfe der STFT können im zugehörigen Spektrogramm nach Gleichung (2.13) nur Frequenzen der Auflösung f_A/K detektiert werden. Eine feinere Detektion der im Signal vorhandenen Frequenzen ist durch die Einbeziehung der Phasenwerte $\theta_\mathrm{STFT}[m,k]$ möglich, auf deren Index ‚STFT' im Folgenden wegen der besseren Lesbarkeit verzichtet wird. Falls sich nicht mehrere Quellensignale in denselben oder benachbarten Frequenzbins überlagern, lassen sich sogar die exakten Frequenzwerte innerhalb von zwei aufeinanderfolgenden Zeitbins bestimmen. Diese genaue und zeitabhängige Frequenz wird als Momentanfrequenz (englisch *instantaneous frequency*, IF) bezeichnet und mithilfe von

$$f_\mathrm{IF}[m,k] = f_\mathrm{STFT}[k] + \frac{\theta\,[m,k] - \hat{\theta}\,[m,k]}{2\pi \cdot \Delta t} \tag{4.7}$$

berechnet [98]. Dabei stellt $f_\mathrm{STFT}[k]$ die Frequenz des k-ten Frequenzbins nach Gleichung (2.13) und

$$\Delta t = \frac{h}{f_\mathrm{A}} \tag{4.8}$$

die Zeitverschiebung zwischen zwei benachbarten Zeitbins dar. Der Phasenschätzwert

$$\hat{\theta}\,[m,k] = \theta\,[m-1,k] + 2\pi \cdot f_\mathrm{STFT}[k] \cdot \Delta t \tag{4.9}$$

resultiert aus der Extrapolation des Phasenwertes $\theta\,[m-1,k]$ des vorhergehenden Zeitschrittes mithilfe der zugehörigen k-ten Binfrequenz. Er repräsentiert den erwarteten Phasenwert, falls das analysierte Signal genau die Frequenz $f_\mathrm{STFT}[k]$ enthält. Bei einem leicht abweichenden

Frequenzwert im Signal entsteht eine Phasendifferenz zwischen tatsächlichem und geschätztem Phasenwert, welche nach Gleichung (4.7) in einen Frequenzfehler umgerechnet werden kann, der zur Berechnung der Momentanfrequenz auf die Binfrequenz addiert wird.

Die Matrix der zeit- und frequenzabhängigen Momentanfrequenzen wäre prinzipiell als eine weitere Zeit-Frequenz-Darstellung für Phaseninformation denkbar, aber ihre Werte steigen mit den Frequenzen der Frequenzbins sehr stark an. Sie sind deshalb nur wenig aussagekräftig, weil von kleinen Phasenabweichungen hervorgerufene Frequenzveränderungen durch diesen Anstieg überlagert werden. Darüber hinaus können sie auch nicht als modifizierte Frequenzwerte der Frequenzbins eingesetzt werden, da sie nicht für den gesamten Signalabschnitt konstant sind und somit nicht für die gesamte Matrix gelten.

Eine Alternative bietet die ausschließliche Betrachtung der Frequenzabweichungen von der jeweiligen Binfrequenz, deren Matrix im Folgenden als Frequenzfehlermatrix

$$\Psi[m,k] = f_{\mathrm{IF}}[m,k] - f_{\mathrm{STFT}}[k] \tag{4.10}$$

$$= \frac{\theta\,[m,k] - (\theta\,[m-1,k] + 2\pi \cdot f_{\mathrm{STFT}}[k] \cdot \Delta t)}{2\pi \cdot \Delta t} \tag{4.11}$$

bezeichnet wird. Ihre Berechnung folgt aus Gleichung (4.7) zur Momentanfrequenz und Gleichung (4.9) des Phasenschätzwertes.

Innerhalb des Zeitintervalls Δt treten für hohe Frequenzen große Phasendifferenzen auf, sodass Mehrdeutigkeiten in der Phasendifferenz entstehen können und die Momentanfrequenz bzw. der Frequenzfehler nicht mehr eindeutig definiert ist. Ab welcher Frequenz das der Fall ist, hängt auch von Δt ab, weswegen oft kleine Werte für die *Hop Size* h gewählt werden. Durch die kleineren Zeitintervalle reduzieren sich die Mehrdeutigkeiten und die Momentanfrequenzen bzw. Frequenzfehler können für mehr Frequenzbins eindeutig berechnet werden. Um trotzdem auch mit Mehrdeutigkeiten der Phasendifferenz umgehen zu können, werden die Phasenwerte im Zähler der Gleichungen (4.7) und (4.11) stets in den Bereich $[-\pi, \pi]$ verschoben. Folglich werden größere Phasendifferenzen und damit auch größere Frequenzfehler nicht berücksichtigt. Da diese aber nur bei hohen Frequenzwerten auftreten und die Instrumententöne

hauptsächlich im tiefen Frequenzbereich liegen, stellt dieses Vorgehen keine erhebliche Einschränkung dar.

Die Zeit-Frequenz-Darstellung einer Frequenzfehlermatrix ist beispielhaft in Abbildung 4.1(f) für das schon in den vorherigen Abschnitten betrachtete Trompetensignal veranschaulicht. Sie hat wenig mit den Darstellungen des STFT-Betrags, der (modifizierten) Gruppenlaufzeiten oder des Produktspektrums gemeinsam, sondern ihre Struktur ähnelt der Darstellung der STFT-Phasenwerte in Abbildung 4.1(e), auch wenn sie Frequenzen angibt und deshalb einen anderen Wertebereich umfasst. In den Bereichen ohne gespielte Trompetentöne weisen sowohl die Frequenzfehlermatrix als auch die STFT-Phasenwerte keine interpretierbaren Strukturen, sondern scheinbar zufällige Werte auf, welche aus den unterschiedlichen Phasenlagen jedes Zeit- und Frequenzbins resultieren. Dagegen bilden sich in den Bereichen der im Signal vorhandenen Tonfrequenzen in beiden Darstellungen regelmäßige Strukturen, anhand derer die genauen Frequenzen identifiziert werden können. Für die reinen Phasenwerte in Abbildung 4.1(e) ist diese Erkennung allerdings schwierig, da die regelmäßigen Strukturen aus der gleichen zeitabhängigen Phase von mehreren nebeneinander liegenden Frequenzbins bestehen. Ein Algorithmus kann aber vor allem lange Linien gleicher Werte detektieren. Solche langen und insbesondere horizontalen Linien entstehen in der Frequenzfehlermatrix um die Bereiche von einer im Signal auftretenden Frequenz, da die Tonfrequenzen in der Regel über mehrere Zeitabschnitte gehalten werden. In Abbildung 4.1(f) sind diese Linien sowohl für die Grund- als auch die Oberschwingungen gut zu erkennen, was die Identifikation der Instrumentenspektren erleichtert.

4.3 Netzarchitekturen zur Instrumentendetektion

Zur Schätzung aktiver Instrumente haben sich, wie in Abschnitt 4.1 beschrieben, künstliche neuronale Netze durchgesetzt. Im Training erlernen sie implizit spezifische Merkmale der charakteristischen Instrumentenspektren, wodurch sie die enthaltenen Musikinstrumente identifizieren können. Die Berücksichtigung komplexer Merkmale erfordert in der Regel eine tiefe Modellarchitektur, die das Risiko eines *Vanishing Gradients* birgt. Um dieses Problem zu umgehen, enthalten die in dieser Arbeit ent-

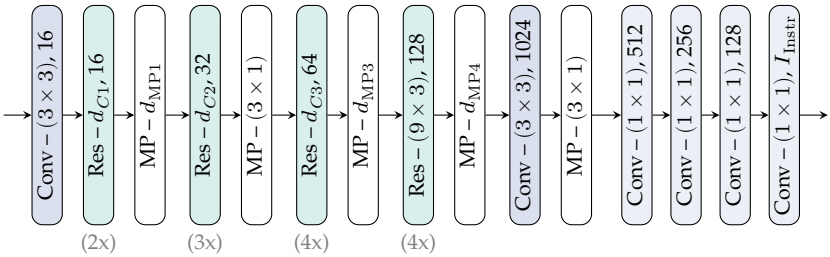


Abbildung 4.2 Schema der Netzarchitekturen ID_ZF_Res und ID_ZF_Deep zur Instrumentendetektion mit Einspeisung von Zeit-Frequenz-Darstellungen, bestehend aus Faltungs- (Conv) und *Max-Pooling*-Schichten (MP) sowie *Residual*-Modulen (Res). Die Zahlen unter den *Residual*-Modulen sind nur in der tiefen Modellvariante ID_ZF_Deep relevant und geben die Anzahl hintereinandergeschalteter *Residual*-Module an.

wickelten Modelle zur Instrumentendetektion mehrere *Residual*-Module, deren Struktur in Abschnitt 2.3.2 vorgestellt wird. Dadurch können auch komplexe Eigenschaften der Musikinstrumente extrahiert werden.

Die zeitabhängige Instrumentendetektion wird in dieser Arbeit sowohl für die Einspeisung von Zeit-Frequenz-Darstellungen als auch für die direkte Einspeisung des Zeitsignals umgesetzt. Für den Fall mit Zeit-Frequenz-Darstellungen werden zwei Netzarchitekturen mit einer tiefen Struktur aus *Residual*-Modulen und Faltungsschichten untersucht. Ihr Aufbau ist in Abbildung 4.2 dargestellt. Der Unterschied zwischen den beiden Modellvarianten liegt in der Anzahl der *Residual*-Module. Das KNN mit den in Abbildung 4.2 dargestellten Blöcken wird im Folgenden als ID_ZF_Res bezeichnet. Darüber hinaus wird auch eine tiefere Modellvariante ID_ZF_Deep untersucht, die anstelle von einem *Residual*-Modul in Abbildung 4.2 jeweils mehrere hintereinandergeschaltete *Residual*-Module enthält. Die Anzahl der hintereinandergeschalteten Module ist unter den abgebildeten *Residual*-Blöcken in grau angegeben.

Alle *Residual*-Module der Modellvariante ID_ZF_Res sind wie in Abbildung 2.5 aufgebaut und besitzen eine Faltungsschicht der Kerngröße (1×1) in der *Skip*-Verbindung, um die Merkmalsdimension des Modulausgangs zu erreichen. Das jeweils erste der hintereinandergeschalteten *Residual*-Module im tiefen Modell ID_ZF_Deep ist identisch aufgebaut. Die übrigen *Residual*-Module von ID_ZF_Deep enthalten keine Faltungs-

schicht in der *Skip*-Verbindung, da sie an Ein- und Ausgang die gleichen Dimensionen besitzen. Am Ende der Modellarchitektur in Abbildung 4.2 erfüllen die Faltungsschichten mit Kerngröße (1×1) die Aufgabe von voll verbundenen Schichten, indem sie die unterschiedlichen Merkmale eines Zeitschritts verknüpfen. Diese Verknüpfung ist durch die Faltung für jeden Zeitschritt gleich, sodass im Vergleich zu voll verbundenen Schichten deutlich weniger Parameter zu trainieren sind und das Modell für jeden Zeitschritt das gleiche Verhalten zeigt. Mit Ausnahme der letzten Schicht wird in allen Faltungsschichten des vorgestellten Modells die ReLU-Aktivierungsfunktion verwendet, um Nichtlinearitäten zu erfassen. In der letzten Faltungsschicht wird dagegen die Sigmoid-Funktion eingesetzt, um einen Schätzwert zwischen 0 und 1 für die Instrumentenaktivität zu erhalten. Tiefe KNN wie die beiden vorgestellten Modelle sind anfällig für Überanpassung, daher wird in allen Faltungsschichten der *Residual*-Module sowie in beiden Faltungsschichten mit Kerngröße (3×3) *Batch*-Normalisierung eingesetzt. Darüber hinaus wird zur Regularisierung *Dropout* mit den Raten 0,5, 0,3 und 0,15 in den drei Faltungsschichten vor der Ausgangsschicht verwendet.

Am Eingang der Modelle ID_ZF_Res und ID_ZF_Deep wird der Absolutbetrag von STFT oder CQT des zu analysierenden Musiksignals eingespeist. Um den Einfluss unterschiedlicher Zeit- und Frequenzauflösungen auf die Instrumentendetektion zu untersuchen, werden in dieser Arbeit verschiedene Dimensionen (K, M) der Eingangsmatrizen mit K Frequenz- und M Zeitbins betrachtet. Darüber hinaus wird die Einspeisung von zusätzlicher Phaseninformation mithilfe der in Abschnitt 4.2 vorgestellten Darstellungen untersucht. Dabei werden die Phasendarstellungen parallel zur Betragsdarstellung der STFT eingespeist, sodass sich die Dimension des Eingangstensors auf $(K, M, 2)$ erweitert. Die Schätzung der zeitabhängigen Instrumentendetektion erfolgt mit einer zeitlichen Auflösung von 92,88 ms, welche durch die Zeitauflösung der untersuchten Zeit-Frequenz-Darstellungen von maximal 46,44 ms sowie durch den Literaturansatz nach Hung und Yang [53] motiviert ist. Aufgrund der im Vergleich zur Eingangsdarstellung leicht reduzierten Zeitauflösung fließen Informationen mehrerer benachbarter Zeitschritte in die Schätzung mit ein, was sie robuster macht. Das Modell schätzt die Aktivität der betrachteten I_{Instr} Musikinstrumente über 30 Zeitsegmente,

Tabelle 4.1 Dimensionen der Faltungskerne d_C und *Max-Pooling*-Bereiche d_{MP} der in Abbildung 4.2 vorgestellten Modellstruktur bei Einspeisung von Zeit-Frequenz-Darstellungen der Dimension (K, M).

(K, M)	d_{C1}	d_{C2}	d_{C3}	d_{MP1}	d_{MP3}	d_{MP4}
$(88, 60)$	(3×3)	(5×3)	(7×3)	(2×1)	(2×1)	(2×2)
$(88, 240)$	(3×3)	(5×3)	(7×3)	(2×2)	(2×2)	(2×2)
$(400, 60)$	(3×3)	(5×3)	(7×3)	(2×1)	(4×1)	(4×2)
$(400, 240)$	(3×3)	(5×3)	(7×3)	(2×2)	(4×2)	(4×2)
$(513, 240)$	(5×1)	(5×1)	(7×1)	(3×2)	(3×2)	(5×2)
$(2049, 60)$	(5×1)	(5×1)	(7×1)	(6×1)	(5×1)	(5×2)
$(2049, 240)$	(5×1)	(5×1)	(7×1)	(6×2)	(5×2)	(5×2)

sodass die boolesche Ausgangsmatrix der zeitabhängigen Instrumentendetektion die Dimension $(30, I_{\mathrm{Instr}})$ besitzt und eine Zeitdauer von 2,79 s umfasst. Folglich muss auch das eingespeiste Musiksignal bzw. seine Zeit-Frequenz-Darstellung eine Länge von 2,79 s aufweisen, weshalb die Darstellung des Gesamtsignals vor der Einspeisung in kürzere Abschnitte dieser Länge unterteilt wird.

Um die geforderte Ausgangsdimension bei Einspeisung der Zeit-Frequenz-Darstellungen zu erhalten, werden die Größen d_{MP} der im *Max-Pooling* zusammengefassten Bereiche in Abhängigkeit der Eingangsdimension (K, M) angepasst. Des Weiteren variieren die Größen der Faltungskerne d_C in den ersten drei *Residual*-Modulen, da sich bei Einspeisung der STFT, welche mehr Frequenzbins K als die CQT aufweist, in Vorversuchen ein stärkerer Fokus auf die Frequenzinformation als hilfreich herausgestellt hat. Daraus ergeben sich die in Tabelle 4.1 aufgelisteten Parameterkombinationen der Modellvariablen aus Abbildung 4.2. Diese Parameter gelten auch für den Fall der parallelen Einspeisung einer zusätzlichen Phasendarstellung mit gleicher Dimension (K, M) wie die Betragsdarstellung. In den *Residual*-Modulen steigt die Größe der Faltungskerne zunehmend an, um komplexe und einen breiteren Frequenzbereich umfassende Merkmale extrahieren zu können. Dabei verdoppelt sich die Anzahl der Merkmale nach Abbildung 4.2 in jeder

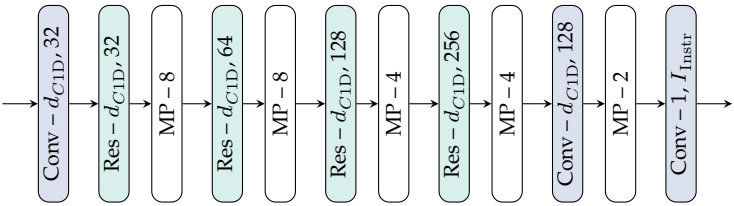


Abbildung 4.3 Schema der Netzarchitektur ID_t_Res zur Instrumentendetektion aus Zeitsignalen, bestehend aus eindimensionalen Faltungs- (Conv) und *Max-Pooling*-Schichten (MP) sowie eindimensionalen *Residual*-Modulen (Res).

Residual-Stufe, damit das Modell ausreichend Information über die im Eingangssignal enthaltenden Instrumente extrahieren kann.

Die zweite Modellarchitektur zur zeitabhängigen Instrumentendetektion besitzt ebenfalls eine Struktur mit *Residual*-Modulen. Am Modelleingang werden aber nur eindimensionale Zeitsignale und keine Zeit-Frequenz-Darstellungen eingespeist, weshalb das KNN ausschließlich eindimensionale Faltungsschichten, *Max-Pooling*-Schichten und *Residual*-Module enthält. Dieses eindimensionale Modell mit direkter Signaleinspeisung, deren Architektur auf der Modellvariante ID_ZF_Res basiert, wird im Folgenden als ID_t_Res bezeichnet. Seine schematische Struktur ist in Abbildung 4.3 dargestellt. Aufgrund der direkten Einspeisung von Zeitsignalabschnitten kommt dieses Modell ohne Signaltransformation in der Vorverarbeitung aus. Des Weiteren sind im Zeitsignal sowohl Betrags- als auch Phaseninformationen der zugehörigen Zeit-Frequenz-Darstellung implizit enthalten, sodass der gesamte Informationsgehalt des Signals eingespeist wird und das KNN daraus selbstständig die relevanten Merkmale zur Instrumentendetektion extrahiert.

Analog zur Struktur ID_ZF_Res sind alle *Residual*-Module des Modells ID_t_Res wie in Abbildung 2.5 aufgebaut, wobei eine eindimensionale Faltungsschicht der Kerngröße 1 in der *Skip*-Verbindung die Daten am Moduleingang an die vorgegebene Anzahl an Merkmalen anpasst. Diese Merkmalsanzahl verdoppelt sich pro *Residual*-Stufe, um zunehmend mehr Information aus dem Musiksignal extrahieren zu können, die zur erfolgreichen Instrumentendetektion beiträgt. Die Ausgangsschicht von ID_t_Res besteht ebenfalls aus einer eindimensionalen Faltungsschicht mit Kerngröße 1. Sie repräsentiert die eindimensionale Version der in

ID_ZF_Res als Ersatz für voll verbundene Schichten genutzten Faltungsschichten mit Kerngröße (1×1). Mehrere solcher Schichten in Reihe würden in diesem Fall zu keiner besseren Verknüpfung der Merkmale führen, weshalb nur eine Ausgangsschicht mit Kerngröße 1 verwendet wird. In allen Faltungsschichten außer der Ausgangsschicht wird die ReLU-Aktivierungsfunktion eingesetzt. Um Schätzwerte zwischen 0 und 1 für die zeitabhängige Instrumentenaktivität zu erhalten, wird in der Ausgangsschicht die Sigmoid-Funktion verwendet. Zur Regularisierung ist *Batch*-Normalisierung in alle Faltungsschichten des Modells ID_t_Res integriert. *Dropout* wird nicht genutzt.

Damit die Ergebnisse aller in dieser Arbeit untersuchten Modelle vergleichbar sind, erfolgt die Schätzung der zeitabhängigen Instrumentenaktivität auch im Falle der direkten Zeitsignaleinspeisung mit einer zeitlichen Auflösung von 92,88 ms. Die Modellarchitektur von ID_t_Res erfordert am Ausgang allerdings eine Potenz der Zahl 2 als Anzahl an geschätzten Zeitsegmenten. Daher wird die Aktivität der I_{Instr} Musikinstrumente über 32 Zeitsegmente geschätzt, sodass die boolesche Ausgangsmatrix der zugehörigen zeitabhängigen Instrumentendetektion die Dimension $(32, I_{\mathrm{Instr}})$ besitzt. Die 32 Segmente umfassen eine Zeitdauer von etwa 2,97 s, die mit der Länge der eingespeisten Zeitsignale übereinstimmen muss. Somit wird das gesamte zu analysierende Musiksignal, das im Falle der direkten Zeitsignaleinspeisung mit 22 050 Hz abgetastet wird, vor der Einspeisung in Abschnitte von 65 536 Werten unterteilt. Mithilfe der *Max-Pooling*-Schichten, deren zusammengefassten Bereiche stets Größen einer Potenz von 2 umfassen, wird die sehr hohe Zeitauflösung des resultierenden Eingangsvektors schrittweise auf die 32 Zeitsegmente am Ausgang reduziert. Als Kerngröße $d_{C1\mathrm{D}}$ der eindimensionalen Faltungsschichten, welche im gesamten Modell gleich ist, können beliebige Dimensionen gewählt werden. In dieser Arbeit werden die Kerngrößen 3, 8 und 16 untersucht, da 3 eine typische Kerngröße in zweidimensionalen Faltungsschichten ist und 8 sowie 16 längere Kerne einer Potenz von 2 darstellen.

4.4 Experimente

Die in Abschnitt 4.3 vorgestellten KNN werden mithilfe eines Datensatzes trainiert und die Güte ihrer Instrumentendetektion anhand der klassischen Klassifikationsmetrik F-Maß evaluiert. Der in diesen Experimenten eingesetzte Datensatz sowie alle wichtigen Parameter und Umsetzungsdetails sind in Abschnitt 4.4.1 beschrieben. Anschließend folgt in den Abschnitten 4.4.2 und 4.4.3 die Analyse der unterschiedlichen Netzarchitekturen sowie der zusätzlichen Phasendarstellungen hinsichtlich der erzielbaren Güte der Instrumentendetektion.

4.4.1 Implementierungsdetails

In überwachten Literaturansätzen zur Instrumentenerkennung werden unterschiedliche Datensätze verwendet. Als Referenzen haben sich hauptsächlich der IRMAS- [5], MedleyDB- [4] und MusicNet-Datensatz [137] etabliert. Da der IRMAS-Datensatz nur die Aktivität des im jeweiligen Musikstück dominanten Musikinstruments bereitstellt, ist er für den betrachteten Fall der polyphonen Instrumentendetektion ungeeignet. Darüber hinaus enthält der MedleyDB-Datensatz viele Musikstücke mit Gesangspassagen, die sich im hier untersuchten Fall von klassischer Instrumentalmusik störend auf die Güte der Instrumentendetektion auswirken würden. Folglich wird das Training und die Evaluation der in dieser Arbeit entwickelten Modelle zur zeitabhängigen Instrumentendetektion mithilfe des MusicNet-Datensatzes durchgeführt. Er umfasst 34 Stunden mit insgesamt 330 frei lizensierten Kammermusikaufnahmen, die mit einer Abtastrate von 44,1 kHz aufgenommen und im ursprünglichen Datensatz in 320 Trainings- und 10 Teststücke unterteilt sind. Um auch schon während des Trainingsprozesses ein Maß für die Detektionsgüte von nicht im Training gelernten Musikstücken zu erhalten, werden die Trainingsstücke hier in einen Trainingsdatensatz mit 286 Stücken und einen Validierungsdatensatz mit 34 Stücken weiter unterteilt.

Obwohl im MusicNet-Datensatz insgesamt elf Musikinstrumente enthalten sind, umfassen die Aufnahmen des Testdatensatzes nur sieben von ihnen. Daher werden in diesem Kapitel nur Modelle und deren Ergebnisse diskutiert, welche die zeitabhängige Aktivität dieser sieben Instrumente Piano, Violine, Viola, Cello, Klarinette, Fagott und Horn schätzen.

Die Musiksignale der anderen vier Instrumente Oboe, Flöte, Cembalo und Kontrabass sind weiterhin Teil der Trainings- und Validierungsdaten, ihre Aktivität wird den KNN aber nicht mitgeteilt. Somit fungieren sie während des Trainings als zusätzliche Störsignale, die das Modell mit $I_{\text{Instr}} = 7$ bei der Instrumentendetektion vernachlässigen muss. Um die Instrumentenaktivitäten jedes Musikstücks in die booleschen Ausgangsmatrizen der Dimension $(30, 7)$ bzw. $(32, 7)$ zu übertragen, welche die *Labels* der Instrumentendetektion darstellen, werden die im MusicNet-Datensatz verfügbaren Informationen aller gespielten Noten den entsprechenden Zeitsegmenten und Musikinstrumenten zugeordnet. Sobald ein Instrument zu einem beliebigen Zeitpunkt innerhalb eines Zeitsegments der durch die Zeitauflösung definierten Länge 92,88 ms eine Note spielt, wird dieses Segment für das entsprechende Instrument auf aktiv gesetzt, was einer 1 in der booleschen Matrix entspricht.

Während des Trainings mit dem MusicNet-Datensatz werden die Parameter aller untersuchten KNN zur zeitabhängigen Instrumentendetektion mithilfe des SGD [37] mit Momentum 0,9 optimiert. Dabei dient die binäre Kreuzentropie (BCE)

$$J_{\text{BCE}} = -\sum_{i=1}^{L} \log\left(\tilde{y}_i\right) \qquad \text{mit} \quad \tilde{y}_i = \begin{cases} \hat{y}_i & y_i = 1 \\ 1 - \hat{y}_i & y_i = 0 \end{cases} \tag{4.12}$$

als Kostenfunktion, die eine Sonderform der in Gleichung (2.28) beschriebenen CE darstellt, bei der die Zielwerte y_i nur die Werte 1 oder 0 annehmen. Folglich passt diese Kostenfunktion sehr gut zur booleschen Ausgangsmatrix der Instrumentendetektion. Die Lernrate ν wird für die ersten fünf Epochen auf 0,1 festgesetzt, anschließend wird sie mit zunehmender Epochenzahl l_{epoch} anhand der Formel

$$\nu = 0{,}1 \cdot 0{,}5^{\left\lfloor \frac{l_{\text{epoch}}+1}{5} \right\rfloor} \tag{4.13}$$

reduziert, damit im Laufe der Optimierung zunehmend feinere Schritte berücksichtigt werden. Alle in diesem Kapitel analysierten KNN werden über insgesamt 50 Epochen trainiert. In der zeitabhängigen Instrumentendetektion werden anschließend die optimierten Parameter der Epoche verwendet, in welcher die mittlere BCE für den im Training unbekannten Validierungsdatensatz am niedrigsten ist.

Als Eingangsdarstellungen werden der Absolutbetrag von STFT oder CQT mit unterschiedlichen Dimensionen (K, M) untersucht. Vor der Einspeisung werden die Betragsdarstellungen jedes gesamten Musikstücks dabei auf ihr Maximum normiert und die Werte in den logarithmischen Bereich überführt, um nicht relevante Lautstärkeeinflüsse zu minimieren. Darüber hinaus wird der Einfluss von zusätzlichen, in Abschnitt 4.2 eingeführten Phasendarstellungen analysiert, die parallel zur Betragsdarstellung der STFT eingespeist werden und daher die gleiche Dimension (K, M) besitzen. Für die Berechnung der STFT und der darauf basierenden Phasendarstellungen werden Analysefenster der Breite 4096 und 1024 verwendet. Diese entsprechen bei der Abtastrate des Datensatzes von 44,1 kHz einer Länge von etwa 92,88 ms bzw. 23,22 ms, was genau mit der Zeitauflösung der Instrumentendetektion bzw. einem Viertel davon übereinstimmt. Während dieser Zeitabschnitte werden die Instrumentensignale jeweils als stationär angenommen. Aus den Fensterbreiten resultieren 2049 bzw. 513 STFT-Frequenzbins. Zur Berechnung der CQT werden die Varianten mit $b = 12$ und $b = 48$ Frequenzbins pro Oktave untersucht, sodass jeder Halbtonschritt entweder ein oder vier Frequenzbins umfasst. Mit einer minimalen Frequenz von 27,5 Hz, welche die Grundfrequenz der Note A_2 repräsentiert, ergeben sich aus den Varianten für b entweder 88 oder 400 CQT-Frequenzbins. Die Anzahl der Zeitbins am Eingang hängt nur von der zeitlichen Schrittweite h ab, da die eingespeiste Zeitdauer auf 2,79 s festgelegt ist. Durch diese *Hop Size* wird die Überlappung zwischen benachbarten Fenstern und somit die in der Zeit-Frequenz-Darstellung enthaltene Redundanz gewählt. Für alle Darstellungen wird sowohl eine *Hop Size* von 512 als auch von 2048 Werten untersucht, woraus sich 240 bzw. 60 Zeitbins ergeben.

Im Gegensatz zur Vorverarbeitung mit Zeit-Frequenz-Transformationen wird die Abtastfrequenz der Zeitsignale für die direkte Einspeisung halbiert. Diese in der Audiosignalverarbeitung übliche neue Abtastfrequenz von 22 050 Hz ermöglicht KNN mit weniger Parametern und reduziert die eingespeiste Datenmenge. Vorversuche mit dem Modell ID_t_Res und eingespeisten Zeitsignalen mit einer Abtastfrequenz von 44,1 kHz zeigen darüber hinaus leicht schlechtere Ergebnisse als im Fall von 22 050 Hz, sodass die reduzierte Abtastfrequenz in allen Experimenten des Modells ID_t_Res verwendet wird.

Aufgrund der Sigmoid-Aktivierungsfunktion in der Ausgangsschicht jedes KNN liegen die Schätzwerte der zeitabhängigen Instrumentendetektion zwischen 0 und 1. Diese reellen Werte sind als Wahrscheinlichkeiten für die Aktivität des jeweiligen Instruments im betrachteten Zeitsegment interpretierbar. Um die Werte in eine Aktivitätsbeschreibung mit binärem Wertebereich zu überführen, werden sie mithilfe eines Schwellenwertes binarisiert. Im vorliegenden Kapitel zur zeitabhängigen Instrumentendetektion wird stets der Schwellenwert 0,5 angewendet, welcher eine robuste Detektion ermöglicht. An die einzelnen Instrumente angepasste Schwellenwerte, die auch in der Literatur vorgeschlagen werden [53], können die Detektionsergebnisse leicht erhöhen [A2]. Daraus können jedoch auch extreme Schwellenwerte resultieren, die zu einer ungewollten, starken Sensitivität für einzelne Instrumente führen.

Zur Bewertung der zeitabhängigen Instrumentendetektion wird hauptsächlich das schon in Gleichung (3.2) vorgestellte F-Maß verwendet. Dabei wird pro betrachtetem Instrument jedes Zeitsegment des Testdatensatzes anhand der geschätzten Aktivität zu den Kategorien korrekt detektiert (TP), falsch detektiert (FP) oder nicht detektiert, obwohl das Instrument aktiv ist (FN), zugeordnet. Genauere Aufschlüsse liefern die ebenfalls in der binären Klassifikation etablierten Größen

$$\mathit{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad \text{und} \qquad \mathit{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}\,, \tag{4.14}$$

aus denen sich wiederum das F-Maß

$$F = \frac{2 \cdot \mathit{Precision} \cdot \mathit{Recall}}{\mathit{Precision} + \mathit{Recall}} \tag{4.15}$$

als harmonisches Mittel der beiden berechnen lässt. Alle Metriken werden für den gesamten Testdatensatz und für jedes Instrument unabhängig berechnet. Als vereinfachtes Maß für die Detektionsgüte dient der Durchschnitt über alle Instrumente.

4.4.2 Vergleich der Netzarchitekturen

Bevor die drei in dieser Arbeit entwickelten Netzarchitekturen aus Abschnitt 4.3 in Abschnitt 4.4.2.2 miteinander verglichen werden, erfolgt eine separate Analyse der einzelnen Architekturen und ihrer veränderbaren Parameter. Im Falle der Modelle ID_ZF_Res und ID_ZF_Deep

Tabelle 4.2 F-Maße des Modells ID_ZF_Res bei Einspeisung von Zeit-Frequenz-Darstellungen mit unterschiedlichen Dimensionen (K, M).

	STFT		CQT	
	$(513, 240)$	$(2049, 240)$	$(88, 240)$	$(400, 240)$
Piano	98,22 %	98,11 %	97,65 %	**98,31 %**
Violine	95,51 %	95,08 %	94,03 %	**95,73 %**
Viola	78,81 %	81,05 %	80,01 %	**81,72 %**
Cello	91,36 %	91,73 %	91,40 %	**92,27 %**
Klarinette	87,75 %	**90,09 %**	87,94 %	89,27 %
Fagott	81,52 %	**83,58 %**	82,76 %	83,54 %
Horn	77,47 %	79,48 %	76,37 %	**80,93 %**
∅	87,23 %	88,44 %	87,17 %	**88,82 %**

wird dabei nur die Einspeisung der Absolutbeträge von STFT und CQT berücksichtigt, da zusätzliche Phasendarstellungen in Abschnitt 4.4.3 untersucht werden.

4.4.2.1 Separate Analyse der Netzarchitekturen

Zunächst wird die zeitabhängige Instrumentendetektion mithilfe des Modells ID_ZF_Res betrachtet. Die F-Maße bei Einspeisung der Betragsdarstellungen von STFT und CQT mit verschiedenen Frequenzauflösungen und der feinen Zeitauflösung mit $h = 512$ sind in Tabelle 4.2 für alle Instrumente des Testdatensatzes gegenübergestellt. Aus den F-Maßen jedes Instruments von mindestens 76 % lässt sich die erfolgreiche zeitabhängige Instrumentendetektion für alle Testinstrumente ableiten. Außerdem wird deutlich, dass die Ergebnisse für Piano und Violine mit über 94 % am besten sind und damit weit über den F-Maßen für Viola, Fagott und Horn liegen. Ein Grund für diese Unterschiede ist die Zusammensetzung des MusicNet-Testdatensatzes, welcher Soloaufnahmen der Instrumente Piano, Violine und Cello enthält. Die anderen vier Instrumente sind im Testdatensatz ausschließlich in Aufnahmen von Trios aktiv, sodass ihre Detektion aufgrund der beiden parallel zu ihnen spielenden Instrumente wesentlich schwieriger ist. Des Weiteren sind

die Instrumentenaktivitäten im gesamten MusicNet-Datensatz ungleich verteilt. In über der Hälfte der Musikstücke spielt unter anderem ein Piano, wohingegen die Instrumente Klarinette, Fagott und Horn nur in weniger als 10 % der Gesamtlänge aktiv sind [137]. Dadurch kann im Training ein Ungleichgewicht entstehen, sodass z. B. das Piano deutlich schneller detektiert wird, weil die Wahrscheinlichkeit für die Aktivität des Pianos im Datensatz höher ist. Folglich lassen sich die Ergebnisse der einzelnen Instrumente nur schwer miteinander vergleichen.

Die Ergebnisse der unterschiedlichen Eingangsdimensionen lassen sich dagegen gut miteinander vergleichen. Anhand der durchschnittlichen F-Maße wird für beide Zeit-Frequenz-Darstellungen deutlich, dass eine höhere Frequenzauflösung am Eingang des Modells ID_ZF_Res die Instrumentendetektion verbessert. Durch die feinere Auflösung der in den Musiksignalen vorhandenen Frequenzen erlernt das KNN genauere Instrumentenspektren und -merkmale, die eine präzisere Detektion der Instrumentenaktivität ermöglichen. Im Falle der CQT mit 400 Frequenzbins sind auch die Einzelergebnisse aller Instrumente besser als mit nur 88 Frequenzbins, sodass das KNN generell von der feineren Auflösung der Signalfrequenzen profitiert. Bei Einspeisung der STFT mit 513 Frequenzbins sind die Ergebnisse für Piano und Violine allerdings leicht besser als im Falle der höheren Frequenzauflösung. Dies zeigt die auch mit niedrigerer Frequenzauflösung sehr gute Instrumentendetektion bei Soloaufnahmen, wohingegen die komplexere Detektion der nur in Trios vorhandenen Instrumente Viola, Klarinette, Fagott und Horn durch die höhere Auflösung klar verbessert wird. Gerade die Instrumente im tieferen Frequenzbereich, wie z. B. Cello, Fagott und Horn, profitieren von einer genaueren Frequenzauflösung der STFT. Insgesamt führen alle in Tabelle 4.2 untersuchten Einspeisungen des Modells ID_ZF_Res zu einer erfolgreichen zeitabhängigen Instrumentendetektion mit einem durchschnittlichen F-Maß über alle Instrumente von mehr als 87 %. Das beste Ergebnis liefert dabei die Betragsdarstellung der CQT mit einer Dimension von $(400, 240)$.

Die tiefere Modellvariante mit Vorverarbeitung, ID_ZF_Deep, besitzt mehr Faltungsschichten und daher mit knapp über 7 Millionen Parametern mehr als doppelt so viele Parameter wie das Modell ID_ZF_Res mit etwa 3 Millionen. Für einen Vergleich der Instrumentendetektion

Tabelle 4.3 Durchschnittliche F-Maße der Modelle ID_ZF_Res und ID_ZF_Deep bei Einspeisung von Zeit-Frequenz-Darstellungen unterschiedlicher Dimensionen (K, M).

	(K, M)	ID_ZF_Res	ID_ZF_Deep
STFT	$(513, 240)$	87,23 %	87,45 %
	$(2049, 60)$	88,39 %	88,05 %
	$(2049, 240)$	**88,44 %**	87,22 %
CQT	$(88, 60)$	86,91 %	85,94 %
	$(88, 240)$	87,17 %	88,05 %
	$(400, 60)$	88,24 %	88,05 %
	$(400, 240)$	**88,82 %**	88,00 %

beider Varianten sind in Tabelle 4.3 die durchschnittlichen F-Maße der Modelle mit allen in dieser Arbeit betrachteten Eingangsdimensionen zusammengestellt. Die dazugehörigen detaillierten F-Maße aller Instrumente sind für beide Modelltypen in separaten Tabellen im Anhang A aufgelistet. Beim Vergleich der durchschnittlichen F-Maße aus Tabelle 4.3 fällt auf, dass die tiefere Modellvariante ID_ZF_Deep in den meisten Fällen eine schlechtere Instrumentendetektion als das Modell ID_ZF_Res erzielt. Dies legt den Schluss nahe, dass die Tiefe von ID_ZF_Res für die zeitabhängige Instrumentendetektion bereits ausreicht. Eine größere Anzahl an Modellparametern erhöht den Ressourcenbedarf sowie die Gefahr von Überanpassung, sodass die Modellvariante ID_ZF_Res mit weniger Parametern für die auf dem MusicNet-Datensatz trainierte Instrumentendetektion geeigneter ist.

Das tiefe Modell ID_ZF_Deep übertrifft das F-Maß des Modelltyps ID_ZF_Res nur bei Einspeisung von Zeit-Frequenz-Darstellungen der Dimensionen $(513, 240)$ (STFT) und $(88, 240)$ (CQT). Im Falle der STFT mit $(513, 240)$ ist die Differenz relativ klein, aber bei Einspeisung des CQT-Absolutbetrags der Dimension $(88, 240)$ ist das F-Maß knapp 1 % höher. Diese Ergebnisse können Ausreißer sein oder auf komplexere Instrumentenmerkmale hindeuten, die das tiefe Modell im Training aufgrund seiner komplexeren Architektur erlernt. Da die besten F-Maße des Modells ID_ZF_Res, die bei Einspeisung von hohen Frequenzauflösungen auftreten, aber deutlich über denen der tiefen Variante ID_ZF_Deep

liegen, wird für die Instrumentendetektion im Folgenden nur noch der Modelltyp ID_ZF_Res als Ansatz mit Einspeisung von Zeit-Frequenz-Darstellungen diskutiert.

Der bereits aus Tabelle 4.2 abgeleitete Zusammenhang zwischen einer feinen Frequenzauflösung und der verbesserten Instrumentendetektion bestätigt sich in den Ergebnissen von Tabelle 4.3 nur teilweise. Für das Modell ID_ZF_Deep sind bei Einspeisung von 240 Zeitbins leicht reduzierte F-Maße für die höhere Frequenzauflösung gegenüber der geringeren Anzahl an Frequenzbins festzustellen. Bei Einspeisung von 60 Zeitbins führt eine höhere Frequenzauflösung dagegen zu verbesserten Ergebnissen, sodass das schlechte Ergebnis von 87,22 % für die STFT-Einspeisung der Größe $(2049, 240)$ als Sonderfall und damit Ausreißer interpretiert werden kann. Als zweite Dimension der Eingangsmatrizen wird in Tabelle 4.3 auch der Einfluss der Zeitauflösung untersucht. Im Vergleich zur Frequenz hat die Anzahl der eingespeisten Zeitbins keinen so großen Einfluss auf die Güte der zeitabhängigen Instrumentendetektion. Eine leichte Verbesserung durch höhere Zeitauflösung ist aber in allen Fällen des Modells ID_ZF_Res zu verzeichnen. Für das tiefere Modell ist der Zusammenhang auch in diesem Fall nicht eindeutig, weil der schon beschriebene Ausreißer mit STFT der Größe $(2049, 240)$ trotz mehr Zeitbins als die STFT der Dimension $(2049, 60)$ ein niedrigeres durchschnittliches F-Maß besitzt. Trotzdem kann für die Einspeisung der Zeit-Frequenz-Darstellungen festgehalten werden, dass eine höhere Auflösung die Instrumentendetektion im Allgemeinen verbessert.

Anstatt in einem Vorverarbeitungsschritt vordefinierte Merkmale mithilfe einer Zeit-Frequenz-Transformation zu extrahieren, ermöglicht die direkte Einspeisung des Zeitsignals in das KNN ein flexibleres Lernen von an die Aufgabe angepassten Merkmalen. Deshalb wird das Modell ID_t_Res mit direkter Signaleinspeisung als Alternative zu den auf Zeit-Frequenz-Darstellungen basierenden Modellen zur Instrumentendetektion untersucht. Die damit erzielten F-Maße aller betrachteten Instrumente sind in Tabelle 4.4 für unterschiedliche Kerngrößen $d_{C1\mathrm{D}}$ der eindimensionalen Faltungsschichten angegeben. Wie auch schon in Zusammenhang mit Tabelle 4.2 erörtert, sind die F-Maße der einzelnen Instrumente aufgrund ihrer Auftrittswahrscheinlichkeiten im Datensatz und der Zusammenstellung des Testdatensatzes sehr unterschiedlich. Um die Unter-

Tabelle 4.4 F-Maße des Modells ID_t_Res mit direkter Einspeisung der Zeitsignale und unterschiedlichen Kerngrößen d_{C1D}. Als Gütefunktionen in der Optimierung werden die BCE, die gewichtete BCE sowie der *Focal Loss* untersucht.

	BCE			gew. BCE	*Focal Loss*
	$d_{C1D} = 3$	$d_{C1D} = 8$	$d_{C1D} = 16$	$d_{C1D} = 8$	$d_{C1D} = 8$
Piano	96,18 %	**97,42 %**	**97,42 %**	97,03 %	97,20 %
Violine	92,46 %	94,14 %	**94,72 %**	94,23 %	94,19 %
Viola	77,55 %	80,69 %	**80,83 %**	79,83 %	77,39 %
Cello	88,86 %	90,77 %	**91,17 %**	91,03 %	89,97 %
Klarin.	84,79 %	**86,48 %**	84,22 %	85,40 %	84,13 %
Fagott	76,32 %	**79,11 %**	78,36 %	**79,98 %**	76,69 %
Horn	71,72 %	**73,24 %**	69,88 %	**74,14 %**	59,82 %
∅	83,98 %	**85,98 %**	85,23 %	85,95 %	82,77 %

schiede der Auftrittshäufigkeiten in den Trainingsdaten auszugleichen, wird im Training des KNN anstelle der in Gleichung (4.12) definierten binären Kreuzentropie (BCE) die gewichtete BCE

$$J_{\text{gewBCE}} = -\sum_{i=1}^{L} \chi_i \cdot \log\left(\tilde{y}_i\right) \tag{4.16}$$

oder der sogenannte *Focal Loss* [78]

$$J_{\text{Focal}} = -\sum_{i=1}^{L} \chi_i \cdot \left(1 - \tilde{y}_i\right)^2 \cdot \log\left(\tilde{y}_i\right) \tag{4.17}$$

als Gütefunktion der Optimierung eingesetzt. Sie verwenden die schon in Gleichung (4.12) eingeführte Variable $\tilde{y}_i$ und gewichten mithilfe des Parameters χ_i, wie stark der Datenpunkt i in die Berechnung einfließt. Der Parameter χ_i wird in dieser Arbeit nach

$$\chi = 1 - \frac{T_{\text{Instr}}}{T_{\text{MusicNet}}} \tag{4.18}$$

für jedes Instrument separat berechnet, wobei die Gesamtzeit T_{Instr} der aktiven Passagen des entsprechenden Instruments im Datensatz auf die

Gesamtdauer T_{MusicNet} aller Musikstücke von MusicNet bezogen wird. Die resultierenden F-Maße mit diesen beiden alternativen Gütefunktionen sind ebenfalls in Tabelle 4.4 angegeben.

Aus den Ergebnissen der untersuchten Kerngrößen wird deutlich, dass die Einbeziehung von ausreichend großen zeitlichen Zusammenhängen eine bessere Instrumentendetektion ermöglicht. Für $d_{C1\mathrm{D}} = 3$ sind diese Zusammenhänge zu klein, sodass die im KNN extrahierten Merkmale nicht alle Informationen beinhalten, welche die Instrumentenerkennung benötigt. Folglich sind die F-Maße aller betrachteten Instrumente schlechter als im Falle von $d_{C1\mathrm{D}} = 8$, was zu einem um 2 % reduzierten durchschnittlichen F-Maß führt. Durch eine noch größere Kerngröße $d_{C1\mathrm{D}} = 16$ kann keine weitere Verbesserung der durchschnittlichen Ergebnisse verzeichnet werden. Dennoch profitiert die Detektion der Streichinstrumente von dieser Kerngröße. Die Ergebnisse der Instrumente Klarinette, Fagott und Horn, die seltener im Trainingsdatensatz vorkommen, sind jedoch merklich schlechter als mit $d_{C1\mathrm{D}} = 8$, sodass die mittlere Kerngröße von 8 für die zeitabhängige Instrumentendetektion mit dem Modell ID_t_Res insgesamt am besten geeignet ist.

Eine weitere Verbesserung der durchschnittlichen Ergebnisse durch die Verwendung der alternativen, gewichteten Gütefunktionen in der Optimierung wird nicht erreicht. Der *Focal Loss* liefert gerade für die wenig im Datensatz vorhandenen Instrumente deutlich schlechtere Resultate als die BCE, was das Gegenteil der ursprünglichen Intention für eine Gütefunktion mit Gewichtung darstellt. Vor allem das Horn wird nur sehr schlecht detektiert, sodass der *Focal Loss* für die untersuchte Instrumentendetektion nicht geeignet ist. Die gewichtete BCE verbessert die F-Maße für Fagott und Horn, welche bisher die Instrumente mit der schlechtesten Detektionsrate waren. Somit entspricht dieses Ergebnis dem gewünschten ausbalancierenden Effekt. Allerdings sind die F-Maße der anderen, im Testdatensatz nur in Trios spielenden Instrumente (Viola und Klarinette) gegenüber der ungewichteten BCE reduziert und der Durchschnitt ist über alle Instrumente vergleichbar, weshalb die gewichteten Gütefunktionen im Folgenden nicht eingesetzt werden.

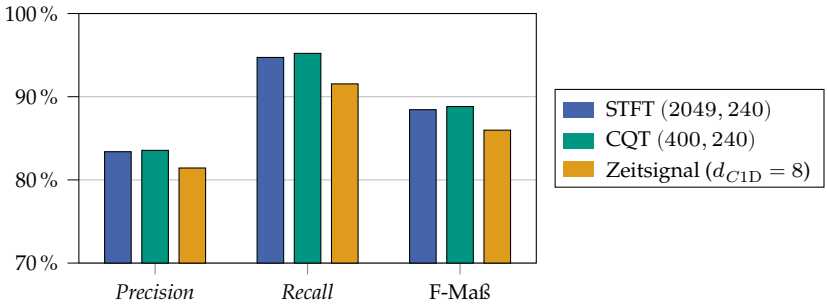


Abbildung 4.4 Vergleich der durchschnittlichen Ergebnisse von den jeweils besten Modellen der Varianten ID_ZF_Res (STFT und CQT) und ID_t_Res (Zeitsignal).

4.4.2.2 Vergleich der Modelle mit und ohne Vorverarbeitung

Obwohl alle hier analysierten Modellarchitekturen, mit Einspeisung von Zeit-Frequenz-Darstellungen und direktem Zeitsignal, den Aktivitätsverlauf der im Musiksignal präsenten Instrumente erfolgreich schätzen, bestätigt der Vergleich ihrer F-Maße den schon in der Literatur für die statische Instrumentendetektion festgestellten Vorteil von Spektrogrammen gegenüber Ende-zu-Ende-Ansätzen [21, 76]. Trotz der Verwendung von tieferen KNN-Architekturen mit *Residual*-Modulen sowie der geforderten Zeitabhängigkeit der Instrumentendetektion ist die Einspeisung von Zeit-Frequenz-Darstellungen der Einspeisung des direkten Zeitsignals auch in dieser Arbeit überlegen. Der Unterschied wird durch die Abbildung 4.4 illustriert, welche die Ergebnisse der besten Modelle für ID_ZF_Res und ID_t_Res anhand ihrer durchschnittlichen *Precision*- und *Recall*-Werte sowie der bereits in den vorangegangenen Tabellen aufgelisteten mittleren F-Maße gegenüberstellt.

Die Modellvariante ID_t_Res weist in allen drei Bewertungsmaßen mindestens 2 % schlechtere Werte als die beiden Modelle des Typs ID_ZF_Res auf, sodass die Überlegenheit der Einspeisung von Zeit-Frequenz-Darstellungen eindeutig belegt wird. Abweichende Vergleichsbedingungen, wie z. B. eine deutlich verschobene Detektionsschwelle von ID_t_Res, können als Grund für das schlechtere F-Maß ausgeschlossen werden, da dann die *Precision*- oder *Recall*-Werte höher als die des Modells ID_ZF_Res sein müssten. Dagegen wird aus den in Tabelle 4.2 und 4.4 angegebenen

F-Maßen der einzelnen Instrumente deutlich, dass die Hauptursache der schlechteren Ergebnisse für ID_t_Res in geringen Detektionsgüten der Instrumente Klarinette, Fagott und Horn liegt. Für diese drei im Trainingsdatensatz seltener vorkommenden Instrumente sind die Ergebnisdifferenzen zwischen ID_ZF_Res und ID_t_Res deutlich größer als für die anderen Instrumente. Eine Erweiterung der Trainingsdaten um Aufnahmen dieser drei Instrumente könnte die Ergebnisse mit direkter Zeiteinspeisung vermutlich etwas verbessern. Doch generell scheinen die wichtigsten Merkmale zur Instrumentenerkennung hauptsächlich auf einer genauen Frequenzverteilung zu beruhen, sodass die Einspeisung von Spektrogrammen vorteilhaft ist. Im Gegensatz zu den direkt eingespeisten Zeitsignalen wird dabei allerdings die Phaseninformation der Zeit-Frequenz-Darstellung weggelassen. Um diese Information ebenfalls in der Instrumentendetektion zu nutzen, wird in Abschnitt 4.4.3 die Einspeisung von zusätzlichen Phasendarstellungen untersucht.

4.4.2.3 Einordnung in Literaturergebnisse

Neben dem Vergleich der in dieser Arbeit entwickelten Modelle zur Instrumentendetektion wird eine Einordnung in die Literatur vorgenommen. Lernende Verfahren, die in den meisten Ansätzen der Instrumentendetektion eingesetzt werden, sind oft nur eingeschränkt miteinander vergleichbar, da sie verschiedene Trainingsdatensätze mit unterschiedlichen Instrumenten und Umfängen verwenden. Darüber hinaus detektieren die meisten Literaturansätze, wie in Abschnitt 4.1 dargestellt, die vorkommenden Instrumente zeitunabhängig. Daher werden hier nur die Ergebnisse der beiden besten Modelle der von Hung und Yang [53] vorgeschlagenen zeitabhängigen Instrumentendetektion betrachtet, die ebenfalls mit dem MusicNet-Datensatz trainiert wurden. Diese KNN besitzen mehrere *Residual*-Module, die aber nur eindimensionale Faltungsschichten enthalten, weshalb ihre Architektur als 1D-*Residual*-Modell (R1D) bezeichnet wird. Basierend auf der etwa 3 s umfassenden CQT-Betragsdarstellung am Eingang schätzen sie die Instrumentenaktivität nur für 28 Zeitsegmente, sodass die Zeitauflösung der Instrumentendetektion mit ungefähr 107 ms etwa 15 % niedriger als in den Schätzungen dieser Arbeit ist. Als zusätzliche Information verwendet ein Modell nach Hung und Yang neben der CQT eine spezielle Darstellung der Tonhö-

Tabelle 4.5 Durchschnittliche Ergebnisse von zwei 1D-*Residual*-Modellen (R1D) nach Hung und Yang [53] und den beiden besten Modellen der Netzarchitektur ID_ZF_Res.

	(K, M)	*Precision*	*Recall*	F-Maß
R1D mit CQT [53]	$(88, 258)$	84,27 %	90,42 %	87,14 %
R1D mit CQT + HSF-5 [53]	$(88, 258)$	**85,44 %**	90,82 %	87,97 %
ID_ZF_Res mit STFT	$(2049, 240)$	83,39 %	94,72 %	88,44 %
ID_ZF_Res mit CQT	$(400, 240)$	83,55 %	**95,21 %**	**88,82 %**

heninformation aller gespielten Noten, die sogenannten *Harmonic Series Features* (HSF), die in der besten Variante Nummer 5 integriert werden. Die durchschnittlichen Ergebnisse dieser Literaturansätze und der beiden besten Modelle dieser Arbeit sind in Tabelle 4.5 angegeben.

Der *Recall* und das F-Maß der beiden in dieser Arbeit vorgestellten Modelle des Typs ID_ZF_Res sind besser als die der Literaturmodelle nach Hung und Yang. Eine Einspeisung von zusätzlichen Informationen über die detektierten Tonhöhen verbessert die Literaturergebnisse, *Recall* und F-Maß bleiben aber unter den Ergebnissen dieser Arbeit, die keine Zusatzinformation miteinbeziehen. Die höhere Detektionsgüte ist zu einem großen Teil auf die feinere Frequenzauflösung der Eingangsdarstellungen zurückzuführen, weil das F-Maß des Modells ID_ZF_Res mit der CQT der Dimension $(88, 240)$ von 87,17 % (s. Tabelle 4.3) fast identisch zu dem des Literaturansatzes R1D mit CQT ist. Aufgrund der 15 % feineren Zeitauflösung der Architektur ID_ZF_Res ist dieses Ergebnis gegenüber dem R1D-Modell aber etwas höher zu bewerten. Darüber hinaus besitzen die Schätzungen mit den Modellen nach ID_ZF_Res eine höhere Robustheit, da sie mithilfe des Schwellenwertes 0,5 binarisiert werden. Für die Modelle mit R1D-Architektur kommen dagegen angepasste Schwellenwerte im Bereich [0,01, 0,99] zum Einsatz, die für jedes Instrument nach dem Training optimiert werden. Dadurch werden Instrumente teilweise nur bei extremen Schätzwerten nahe 0 oder 1 als nicht aktiv bzw. aktiv detektiert, was die Robustheit der Detektion senkt. Ein weiterer Effekt dieser angepassten Schwellenwerte sind die hohen *Precision*- und die vergleichsweise niedrigen *Recall*-Werte. Für die Weiterverwendung der geschätzten Instrumentenaktivitäten ist in vielen MIR-

Aufgaben ein hoher *Recall* positiv, da in diesem Fall mehr vorkommende Instrumente richtig detektiert werden. Die vorgestellte zeitabhängige Instrumentendetektion des Typs ID_ZF_Res ist somit sowohl aufgrund ihrer höheren Detektionsgüte als auch wegen des höheren *Recalls* besser für die Einspeisung in ein nachgeschaltetes Separationssystem geeignet als die Literaturansätze.

4.4.3 Einbeziehung von Phaseninformation

Netzarchitekturen auf Basis von Zeit-Frequenz-Darstellungen sind nach Abschnitt 4.4.2.2 vergleichbaren Modellen mit direkter Einspeisung von Zeitsignalen überlegen. Dabei wird die Betragsdarstellung der STFT oder CQT in das KNN eingespeist und somit ihre Phaseninformation vernachlässigt. Um diese Information ebenfalls in der Instrumentendetektion berücksichtigen zu können, wird neben der verwendeten Betragsdarstellung eine der in Abschnitt 4.2 vorgestellten Phasenrepräsentationen eingespeist, sodass die Eingangsdimension auf $(K, M, 2)$ wächst. Die Phaseninformation verfeinert hauptsächlich den tiefen Frequenzbereich, weswegen die in den unteren Frequenzen bereits sehr gut aufgelöste CQT wenig von einer zusätzlichen Phasendarstellung profitieren sollte. Folglich werden in dieser Arbeit nur Phasendarstellungen analysiert, die auf der STFT basieren.

Gerade für die STFT-Einspeisung mit niedrigerer Frequenzauflösung bietet sich die zusätzliche Phaseninformation an, um feiner aufgelöste spektrale Merkmale extrahieren zu können. Für diesen Fall sind in Tabelle 4.6 die F-Maße der zeitabhängigen Instrumentendetektion mit und ohne Einspeisung der in Abschnitt 4.2 beschriebenen Phasendarstellungen der modifizierten Gruppenlaufzeit (ModGD), des Produktspektrums (PS) und der Frequenzfehlermatrix Ψ aufgelistet. Am durchschnittlichen F-Maß, das bei jeder Phasenrepräsentation gegenüber der reinen Betragseinspeisung verbessert ist, kann die generelle Nützlichkeit der zusätzlichen Phaseninformation abgelesen werden. Doch nicht nur der Durchschnitt, sondern auch die Detektion fast aller Instrumente profitiert von der Zusatzeinspeisung. Speziell die Ergebnisse der in den Testdaten nur als Trio vorkommenden Viola, Klarinette, Fagott und Horn sind bei allen hier untersuchten Phasendarstellungen besser als bei der reinen Einspeisung des STFT-Betrags.

Tabelle 4.6 F-Maße des Modells ID_ZF_Res bei Einspeisung von Betrags- (STFT) und Phasendarstellungen (ModGD, PS und Ψ) der Dimension $(513, 240)$.

	STFT	STFT + ModGD	STFT + PS	STFT + Ψ
Piano	98,22 %	98,04 %	97,97 %	**98,33 %**
Violine	95,51 %	95,66 %	95,65 %	**95,89 %**
Viola	78,81 %	80,76 %	**83,03 %**	82,40 %
Cello	91,36 %	91,59 %	**92,15 %**	91,36 %
Klarinette	87,75 %	**88,43 %**	88,09 %	88,14 %
Fagott	81,52 %	**82,79 %**	82,64 %	81,82 %
Horn	77,47 %	77,86 %	78,60 %	**78,93 %**
∅	87,23 %	87,88 %	**88,30 %**	88,12 %

Insgesamt liefert die Hinzunahme des Produktspektrums die besten Ergebnisse der zeitabhängigen Instrumentendetektion für Einspeisungen mit einer Matrixdimension von $(513, 240)$, gefolgt von der Frequenzfehlermatrix und der modifizierten Gruppenlaufzeit. Dies bestätigt den visuellen Eindruck aus Abschnitt 4.2.2, dass die Darstellung des Produktspektrums gegenüber der modifizierten Gruppenlaufzeit klarer ist und weniger Artefakte besitzt. Da sowohl das der STFT-Betragsdarstellung ähnlich sehende Produktspektrum als auch die der STFT-Phasendarstellung ähnlich sehende Frequenzfehlermatrix zu guten Ergebnissen der Instrumentendetektion führen, kann für die zusätzliche Phasendarstellung kein Vorteil in der Ähnlichkeit zur Darstellung von STFT-Betrags- oder STFT-Phasenwerten festgestellt werden.

Bei Einspeisung von STFT-Betragsdarstellungen mit einer größeren Anzahl an Frequenzbins ist die Frequenzauflösung der Darstellung bereits hoch, sodass die zusätzliche Einspeisung von Phaseninformation keine große Verbesserung für die Instrumentendetektion bringt. Dies wird in Abbildung 4.5 anhand der Detektionsergebnisse für die schon in Abschnitt 4.4.2 untersuchten Dimensionen der STFT verdeutlicht. Die zugehörigen Werte der dargestellten F-Maße sind in Tabelle A.1 des Anhangs aufgelistet. Für 2049 Frequenzbins führt die parallele Einspeisung der modifizierten Gruppenlaufzeit und der Frequenzfehlermatrix sogar

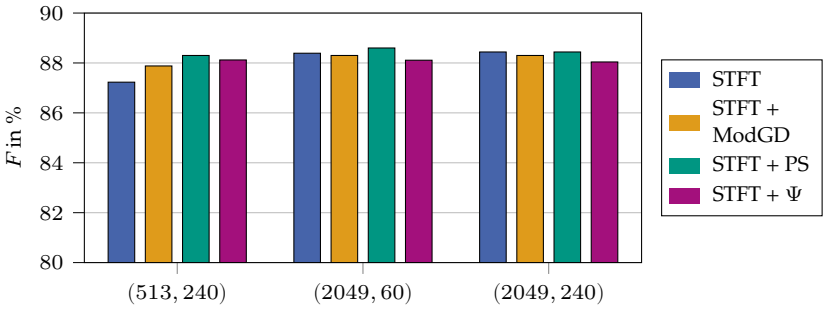


Abbildung 4.5 Vergleich des durchschnittlichen F-Maßes des Modelltyps ID_ZF_Res bei Einspeisung von Betrags- (STFT) und Phasendarstellungen (ModGD, PS und Ψ) mit unterschiedlichen Dimensionen (K, M).

zu schlechteren Ergebnissen als ohne Phasenrepräsentation. Nur die schon bei einer niedrigeren Frequenzauflösung beste Phasendarstellung, das Produktspektrum, liefert auch im Falle mit 2049 Frequenzbins ein im Vergleich zur reinen Betragseinspeisung immer mindestens gleich gutes durchschnittliches F-Maß. Die Verbesserung durch das Produktspektrum ist bei feinerer Frequenzauflösung aber nur sehr gering, wohingegen die Berechnung und die zusätzliche Eingangsmatrix mehr Ressourcen benötigen. Daher ist der Einsatz einer parallelen Phasendarstellung nur bei kleineren Dimensionen der eingespeisten STFT sinnvoll. Dann hilft sie, vor allem im tiefen Frequenzbereich genauere spektrale Merkmale zu extrahieren. Bei ausreichend guter Frequenzauflösung in der Betragsdarstellung bringt die zusätzliche Phaseninformation dagegen keinen Mehrwert und kann weggelassen werden.

Alle diskutierten Ergebnisse mit Phasendarstellungen verwenden die Netzarchitektur ID_ZF_Res, welche nach Abschnitt 4.4.2.1 für reine Betragsdarstellungen besser als die tiefere Architektur ID_ZF_Deep abschneidet und weniger Ressourcen benötigt. Im Allgemeinen werden sie durch die Ergebnisse mit Phasendarstellungen für die Modelle des tiefen Typs ID_ZF_Deep bestätigt, welche in Tabelle A.2 zusammengestellt sind. So liefert die zusätzliche Einspeisung des Produktspektrums die besten Detektionsergebnisse für die Eingangsdimension $(513, 240)$. Für die Modelle mit höherer Frequenzauflösung liefern die Ansätze der Netzar-

chitektur ID_ZF_Deep meist vergleichbare durchschnittliche Ergebnisse wie die in Abbildung 4.5 illustrierten. Ein Ausreißer ist die Einspeisung von STFT-Betrag und Frequenzfehlermatrix der Dimension $(2049, 60)$, deren Gesamtergebnis von 88,74 % besser ist als alle anderen mit STFT-Einspeisung, aber trotzdem noch unter dem besten CQT-Ergebnis des Modelltyps ID_ZF_Res liegt. Insgesamt bestätigt sich aber in den meisten Fällen, dass die durchschnittlichen Ergebnisse des Typs ID_ZF_Deep etwas schlechter als die Ergebnisse der Netzarchitektur ID_ZF_Res sind, daher werden sie hier nicht genauer betrachtet.

5 Separation polyphoner Ensemble-Aufnahmen

Eine gemeinsame Aufnahme von mehreren gleichzeitig gespielten Tönen unterschiedlicher Musikinstrumente lässt sich in der Regel nur noch in ihrer Gesamtheit bearbeiten. Deshalb werden die Signale der einzelnen Instrumente in Studioaufnahmen auf separaten Spuren isoliert aufgenommen. Um aus einem digitalen Gesamtsignal nachträglich dennoch getrennte Instrumentensignale extrahieren zu können, ist Vorwissen über die Instrumentencharakteristiken oder ihre gespielten Töne notwendig. Dieses Vorwissen wird in den aktuell besten Separationsansätzen, die alle auf KNN beruhen, vorwiegend implizit durch die Trainingsdaten integriert, aus denen die charakteristischen Eigenschaften der betrachteten Musikinstrumente erlernt werden. Abschnitt 5.1 gibt einen Überblick über die wichtigsten Ansätze zur Separation von Musiksignalen.

In der Literatur wird oft die Trennung der vier Quellen Gesang, Bass, Schlagzeug und Rest betrachtet. Ihre Charakteristiken sind sehr verschieden, sodass die Separation leichter ist als z. B. bei polyphonen Musikaufnahmen von Instrumenten, die in einem ähnlichen Frequenzbereich spielen. Diesen schwierigeren Fall von sich überlagernden Musikinstrumenten untersucht die vorliegende Arbeit für monaurale Aufnahmen, d. h. über nur ein Mikrofon aufgenommene Musik. Das dazu verwendete Separationsmodell wird in Abschnitt 5.2 vorgestellt. Es wird anhand eines Datensatzes aus instrumentalen Kammermusikaufnahmen mit verschiedenen Ensemble-Besetzungen trainiert, dessen Zusammensetzung in Abschnitt 5.3 beschrieben ist. Die Quellentrennung kann entweder als gemeinsame Aufgabe oder in Form von mehreren Teilaufgaben, welche die Extraktion jeweils eines Instrumentensignals realisieren, umgesetzt werden. Bei einer gemeinsamen Optimierung der Teilaufgaben während des Trainings handelt es sich um einen Multi-Task-Ansatz, dessen Eignung für die Separation polyphoner Ensemble-Aufnahmen in Ab-

schnitt 5.4 untersucht wird. Um die Separationsergebnisse zu verbessern, werden zusätzlich zum monauralen Gesamtsignal Informationen über die zeitabhängigen Instrumentenaktivitäten eingespeist. Ihre Auswirkungen werden in Abschnitt 5.5 sowohl für gemeinsame Separationsmodelle als auch für Einzelmodelle jedes Instruments analysiert. Diese Analysen basieren auf den aus dem Datensatz bekannten und in der Simulation mit zufälligen Fehlern belegten Instrumentenaktivitäten. Da die wahren Aktivitäten in der realen Anwendung jedoch nicht bekannt sind, erfolgt in Abschnitt 5.6 die Verifizierung der simulativ erzeugten Separationsergebnisse anhand von realen Zusatzinformationen des in Kapitel 4 vorgestellten Ansatzes zur zeitabhängigen Instrumentendetektion.

Große Teile der in diesem Kapitel vorgestellten Separation mit Einspeisung von Zusatzinformation über die zeitabhängige Instrumentenaktivität wurden bereits in [A4] veröffentlicht.

5.1 Quellentrennung von Musiksignalen

Die Trennung von monauralen Musiksignalen wurde in der Literatur anfangs als eine Anwendung der blinden Quellentrennung (englisch *blind source separation*, BSS) von Audiosignalen verstanden. Bei dieser Sonderform der Quellentrennung ist kein Vorwissen über die Signale notwendig, weshalb sie für unterschiedliche Signalquellen einsetzbar ist. Ein klassischer BSS-Algorithmus ist die Analyse unabhängiger Signalkomponenten (englisch *independent component analysis*, ICA), welche das Eingangssignal in durch Optimierung identifizierte, unabhängige Komponenten separiert. Einfache Musiksignale aus zwei Quellen, wie z. B. Flöte und Bass, können damit auf Basis einer STFT- [3] oder einer Wavelet-Darstellung [97] getrennt werden. Musiksignalspektren besitzen ausschließlich positive Werte, sind aufgrund ihrer Zusammensetzung aus Grund- und Oberschwingungen in der Regel aber nicht unabhängig. Deshalb ist die eng mit der ICA verwandte, nichtnegative Matrixfaktorisierung (englisch *nonnegative matrix factorisation*, NMF) für die Quellentrennung von Musiksignalen besser geeignet. Sie optimiert die Zerlegung der Eingangsmatrix in zwei Matrizen kleinerer Dimension, die miteinander multipliziert wieder die Eingangsmatrix ergeben. Dabei kann der Inhalt der ersten resultierenden Matrix als Bibliothek der vorkommen-

den Notenspektren und die zweite Matrix als zeitlicher Aktivitätsverlauf der gespielten Notenwerte interpretiert werden. Der erste Ansatz für Klaviersignale [127] wurde durch zahlreiche Weiterentwicklungen wie z. B. die Einbeziehung von zeitlicher Kontinuität [147] oder die Nutzung des Active-Set-Algorithmus in der Optimierung [148] verbessert und auf andere Anwendungsgebiete übertragen. Ein Überblick über die zahlreichen NMF-basierten Ansätze zur Musiksignalseparation geben Févotte et al. [32]. Auch wenn der reine NMF-Algorithmus ein Verfahren zur blinden Quellentrennung ist, integrieren verbesserte Ansätze häufig Vorwissen über die zu trennenden Signale, wodurch die Separation nicht mehr als blind bezeichnet werden kann. Dieser Trend ist im Überblick von Vincent et al. [144] ausführlich beschrieben und setzt sich durch die Verwendung von überwachten KNN in neueren Ansätzen fort, welche in den Abschnitten 5.1.1 und 5.1.2 vorgestellt werden.

Zur allgemeinen Bewertung der getrennten Quellensignale ist ein objektives Maß unabdingbar, das sowohl den Grad der Trennung von den übrigen Quellen als auch die Güte des geschätzten Signals miteinbezieht. Solch ein Bewertungsmaß liefert das analog zum SNR auf Verhältnissen von definierten Signalenergien beruhende *Source-to-Distortion Ratio* (SDR) [143]. Darüber hinaus kann die Güte der Trennung mithilfe des *Source-to-Interferences Ratio* (SIR) und der Einfluss von Artefakten über das *Source-to-Artifacts Ratio* (SAR) separat bewertet werden. Eine Erweiterung dieses Ansatzes sind die subjektiven Bewertungsmaße nach Emiya et al. [28], die Parameter menschlicher Wahrnehmung miteinbeziehen. Obwohl das SDR sehr verbreitet ist, berücksichtigt es in der Berechnung oft nur einzelne Frequenzbereiche des zu analysierenden Signals. Dieses Problem wird durch die Modifikation des skalierungsinvarianten SDR (SI-SDR) [72] behoben. Analog zur Definition des SDR teilt es das geschätzte Quellensignal

$$\hat{s} = e_{\text{target}} + e_{\text{interf}} + e_{\text{artif}} \tag{5.1}$$

in den gewünschten Signalanteil e_{target} der Quelle, die summierten Signalanteile e_{interf} anderer Quellen sowie die Artefakte e_{artif} auf. Daraus lässt sich dann das gewünschte Maß

$$\text{SI-SDR} = 10 \log_{10} \left(\frac{\|e_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{artif}}\|^2} \right) \tag{5.2}$$

sowie die dazugehörigen Teilmaße

$$\text{SI-SIR} = 10\log_{10}\left(\frac{\|e_{\text{target}}\|^2}{\|e_{\text{interf}}\|^2}\right) \quad \text{und} \quad \text{SI-SAR} = 10\log_{10}\left(\frac{\|e_{\text{target}}\|^2}{\|e_{\text{artif}}\|^2}\right) \tag{5.3}$$

berechnen, die in dieser Arbeit zur Bewertung der Ergebnisse aller Separationsverfahren verwendet werden.

5.1.1 KNN-basierte Separationsmodelle

Durch die Einführung von Methoden des *Deep Learnings* konnte die Quellentrennung von Musiksignalen im Allgemeinen, vor allem aber auch die Trennung von monauralen Signalen deutlich verbessert werden. Die Separation profitiert dabei von nichtlinearen Zusammenhängen und Instrumentenmerkmalen, die das Separationsmodell während des überwachten Trainingsprozesses implizit aus den Trainingsdaten erlernt. Mehrere Ansätze mit unterschiedlichen Netzarchitekturen, wie einer klassischen CNN-Struktur [10], einem *Denoising* Autoencoder [38] oder einem *Variational* Autoencoder [102], welcher mithilfe einer Zufallsvariable in der latenten Schicht des Autoencoders eine zusätzliche Variabilität der Instrumentenspektren modelliert, liefern klar bessere Separationsergebnisse als klassische Ansätze zur Quellentrennung wie die NMF. Dies verdeutlicht die Überlegenheit der KNN in diesem Bereich. Eine weitere Verbesserung der Quellentrennung kann durch die Fusion eines klassischen KNN aus voll verbundenen Schichten und eines rekurrenten KNN erzielt werden [139], wobei das rekurrente Netz aus bidirektionalen LSTM-Modulen aufgebaut ist.

Um die verschiedenen Ansätze vergleichen zu können, hat sich die Separation der vier Komponenten Gesang, Schlagzeug, Bass und Rest etabliert. Sehr oft erfolgt der Vergleich auf Basis des weit verbreiteten MUSDB18-Datensatzes [108]. Spezifischere Szenarios wie die reine Extraktion einer Quelle, oft der Gesangsstimme [56], oder die Trennung von Melodie und Begleitung [109] verwenden sehr ähnliche Ansätze, werden hier aber aufgrund ihrer einfacheren Aufgabe nicht behandelt. Darüber hinaus werden keine Separationsansätze betrachtet, die auf eine definierte Nachbearbeitung wie die Neuabmischung der Musiksignalan-

teile ausgerichtet sind [151]. Sie schränken das breite Anwendungsfeld der Quellentrennung zu stark ein.

Für die Separation der vier oben beschriebenen Komponenten haben sich zwei Grundkonzepte entwickelt. Ein Konzept basiert auf der Zeit-Frequenz-Darstellung des zu separierenden Eingangssignals, häufig der STFT, und schätzt am Modellausgang pro Quelle eine Matrix im reellen Wertebereich zwischen 0 bis 1. Diese sogenannte Maske wird entweder direkt vom KNN geschätzt oder mithilfe eines Wiener Filters aus den vom KNN geschätzten Spektrogrammen der zu separierenden Quellen berechnet. Anschließend werden die Masken mit der Zeit-Frequenz-Transformierten des ursprünglichen Gesamtsignals multipliziert und die resultierenden Matrizen in den Zeitbereich rücktransformiert, um die getrennten Quellensignale zu erhalten. Referenzsysteme dieses ersten Konzepts sind der hauptsächlich aus bidirektionalen LSTM-Modulen bestehende Ansatz Open-Unmix [132], das auf Geschwindigkeit optimierte System Spleeter [46] und das D3Net [136], das mehrere parallele Teilnetze für unterschiedliche Frequenzbänder besitzt, deren Merkmale schließlich in einem gemeinsamen Block fusioniert werden. Im Gegensatz zu den anderen beiden genannten Ansätzen trainiert Open-Unmix jeweils ein Modell pro zu trennender Quelle, sodass es flexibel auf andere Besetzungen erweitert werden kann. Da die Masken normalerweise reelle Werte besitzen, wird die Phaseninformation der Zeit-Frequenz-Transformierten in der Schätzung vernachlässigt und für die getrennten Quellensignale ohne Änderung aus dem Eingangssignal übernommen. Somit ist die erreichbare Separationsqualität beschränkt und kann nur durch die Schätzung von komplexen Masken [69] erhöht werden. Luo und Yu [83] kombinieren die Schätzung von komplexen Masken mit der parallelen Verarbeitung von Subbändern verschiedener Frequenzbereiche der eingespeisten STFT. Dadurch erzielt ihre *Residual*-RNN-Architektur mit bidirektionalen LSTM-Modulen die aktuell besten Separationsergebnisse bei ausschließlichem Training mit dem MUSDB18-Datensatz.

Das zweite Grundkonzept zur Quellentrennung speist die Zeitsignale direkt in das KNN ein und schätzt am Ausgang keine Masken, sondern unmittelbar die getrennten Quellensignale. Folglich wird implizit auch die Phaseninformation miteinbezogen und das Modell extrahiert die für die vorgegebene Aufgabe wichtigen Merkmale. Aufgrund der vielen

Datenpunkte des eingespeisten Zeitsignals haben die Modelle dieses Konzepts deutlich mehr Parameter als die des ersten Grundkonzepts. Die erste erfolgreiche Umsetzung des Konzepts im Zeitbereich ist das auf einer UNet-Architektur basierende Wave-U-Net [131]. Ein weiterer Ansatz ist das aus der Sprachsignalverarbeitung auf Musiksignale übertragene Conv-Tasnet [18], das in der ersten Schicht angepasste Filter mit im Training optimierten Parametern besitzt, deren Merkmale als Signalbibliothek dienen. Am Modellausgang werden diese Merkmale dann mit in weiteren Schichten extrahierten Gewichten zu den geschätzten Quellensignalen zusammengefügt. Der beste Ansatz des zweiten Grundkonzepts mit direkter Zeitsignaleinspeisung ist Demucs [18], welcher das Wave-U-Net durch Elemente der Signalsynthese verbessert. So nutzt Demucs eine sehr große Anzahl an Merkmalen, die im Decoder mithilfe von transponierten Faltungsschichten wieder auf die ursprüngliche Zeitauflösung gebracht werden, und verwendet keine *Batch*-Normalisierung. Darüber hinaus sorgen LSTM-Module für die Einbeziehung von zeitlichen Zusammenhängen und gesteuerte Aktivierungsfunktionen (englisch *gated linear unit*, GLU) für merkmalsabhängige Nichtlinearitäten. Unter Verwendung des MUSDB18-Datensatzes war der Demucs-Ansatz über die letzten Jahre das führende System zur Trennung von Gesang, Schlagzeug, Bass und Rest. Insgesamt erreichen beide Grundkonzepte der Quellentrennung, d. h. die Separation auf Basis einer Zeit-Frequenz-Darstellung sowie die direkte Schätzung der Quellensignale aus dem Zeitsignal, aber vergleichbare Ergebnisse.

Neueste Separationssysteme wie das KUIELab-MDX-Net [65] oder Hybrid Demucs [17] schlagen die Fusion von zwei parallelen KNN vor, wobei ein Teilmodell die Zeit-Frequenz-Darstellung des Eingangssignals und das andere Teilmodell das Zeitsignal separiert. Auch wenn die Quellentrennung durch diese Fusion der Grundkonzepte verbessert wird, weisen die Ansätze große Architekturen mit vielen Parametern auf, sodass sehr viele Trainingsdaten benötigt werden. Der Bedarf an großen Trainingsdatensätzen wird in neuen Ansätzen durch die zunehmende Verwendung von Transformer- bzw. *Attention*-Strukturen [141] gesteigert, welche ebenfalls sehr große Mengen an Trainingsdaten benötigen. Ihre Überlegenheit gegenüber bisherigen Ansätzen zur Trennung von monauralen Musik- und Störsignalen, d. h. zur Entrauschung von Mu-

sikaufnahmen, konnte anhand von synthetischen Musiksignalen gezeigt werden [152]. Für die Separation von mehreren Quellen wie den vier oben genannten Komponenten Gesang, Schlagwerk, Bass und Rest erreicht das auf der diskreten Kosinustransformation basierende, effiziente MultiResUNet-*Framework* [123] mit speziellen Multi-*Residual*-Blöcken und *Attention*-Modulen vergleichbare Ergebnisse zur Literatur. Übertroffen werden sie vom System Hybrid Transformer Demucs [112], das eine Weiterentwicklung des Fusionsansatzes Hybrid Demucs darstellt und in der Mitte seiner Architektur die Extraktion von domänenübergreifenden *Attention*-Merkmalen ermöglicht. Damit dieser Ansatz ausreichend trainiert werden kann, sind allerdings sehr viele Musikaufnahmen nötig.

5.1.2 Separation mit Einspeisung von Vorwissen

Um die Quellentrennung weiter zu verbessern, werden Separationssystemen Zusatzinformationen zur Verfügung gestellt. Eine Möglichkeit der Zusatzinformation ist die Partitur, die alle Notenwerte der spielenden Musikinstrumente enthält. Sie ermöglicht unter anderem eine verbesserte Separation von synthetisch erzeugter, klassischer Musik mit vier Instrumenten [94]. Aufgrund der sehr eingeschränkten Klangvariabilität von synthetischen Instrumentensignalen sind diese zwar relativ einfach mithilfe von KNN zu trennen, die Übertragung auf andere Klänge ist aber problematisch. Folglich können sie zur Optimierung der Syntheseparameter einzelner Instrumente hilfreich sein, bei welcher der Fehler zwischen dem Eingangsmix und der Summe aller resynthetisierten Instrumentensignale minimiert wird [59]. Für die Quellentrennung eignen sich synthetische Signale aber wegen der geringen Generalisierungsfähigkeit nur schlecht. Darüber hinaus bleibt beim Ansatz mit eingespeister Partitur der Nachteil bestehen, dass die zeitsynchronisierte Partitur des gespielten Stücks bekannt sein muss. Diese Vorgabe kann durch die Fusion von Musikseparation und Transkription [84] vermieden werden, in der die separierten Instrumentenspuren und die Partitur in einem Multi-Task-Ansatz gemeinsam geschätzt werden. Ein weiterer Multi-Task-Ansatz zur Quellentrennung ist die gemeinsame Schätzung von Instrumentenspuren und Instrumentenaktivität [52]. Beide Ansätze zeigen, dass jede Einzelaufgabe von der anderen profitieren kann. Da die verschiedenen Aufgaben aber häufig auch viele unterschiedliche Merk-

male benötigen, sind die resultierenden KNN sehr groß und brauchen große Trainingsdatensätze für eine erfolgreiche Parameteroptimierung.

Für die Integration von Zusatzinformationen über aktive Instrumente wird meistens eine bedingte Quellentrennung verwendet, bei der ein binärer Vektor eingespeist wird, welcher die vorab bekannte Präsenz der berücksichtigten Instrumente im betrachteten Musikstück angibt. Dieser Vektor kann beispielsweise durch Multiplikation mit der latenten Repräsentation integriert werden [126]. Häufig separiert die bedingte Quellentrennung nicht alle Instrumentensignale gleichzeitig, sondern extrahiert nacheinander immer nur eine Quelle. Dabei wird die Extraktion stets auf dem gleichen Separationsmodell durchgeführt, das mithilfe eines Einheitsvektors als zweitem Modelleingang gesteuert bzw. über das zu trennende Instrument informiert wird. Die Integration des Einheitsvektors erfolgt mithilfe eines zusätzlichen neuronalen Kontrollnetzes und speziellen Schichten im Separationsmodell, wie z. B. die *Feature-wise Linear Modulation* (FiLM) Schicht [91] oder die *Latent Source Attentive Frequency Transformation* (LaSAFT) [13]. Neben einem Einheitsvektor sind auch andere Eingangsdaten für das Kontrollnetz möglich, beispielsweise eine ungestörte Beispielaufnahme des zu trennenden Instruments [85] oder eine Videosequenz [125], um das zu trennende Instrument zu charakterisieren. Darüber hinaus kann über den bedingten Ansatz sogar die MIR-Aufgabe gewählt werden, die vom System ausgeführt werden soll. So ermöglicht das Modell von Lin et al. [77] je nach Kontrollvektor eine Separation, Transkription oder Synthese. Dies wird in der *Encoder-Decoder*-Architektur des Modells durch eine Entkopplung von Tonhöhe und Klang realisiert. Des Weiteren kann das zu trennende Instrument anhand eines Audiobeispiels vorgegeben werden, das über ein Kontrollnetz in den latenten Raum des Separationsmodells eingespeist wird.

Die meisten Ansätze zur bedingten Quellentrennung nutzen keine zeitabhängigen, sondern nur statische Zusatzinformationen wie das zu trennende Instrument oder seine Klangcharakteristik. Gerade für die Separation zeitabhängiger Musiksignale kann zeitabhängiges Zusatzwissen aber von Vorteil sein. Bei bekanntem Liedtext führt eine Einspeisung der zeitabhängigen Phoneme, also der kleinsten bedeutungsunterscheidenden lautlichen Einheiten einer Sprache, zu Verbesserungen der Gesangsextraktion [90]. Die Matrix der zeitabhängigen Phoneme wird

dabei aus den jeweiligen Liedtexten erstellt. Für die Quellentrennung einer Musikaufnahme mit mehreren Instrumenten kann Vorwissen über die zeitabhängige Instrumentenaktivität durch eine zeitliche Zerlegung der Gesamtaufnahme durch den Anwender ausgenutzt werden [9]. Das ermöglicht die Anpassung des Separationsmodells auf den jeweiligen Musikabschnitt mit vorgegebener Besetzung.

5.2 Separationsmodell

In dieser Arbeit wird die automatische Separation von mehrstimmigen Ensemble-Aufnahmen mit und ohne zusätzliche Einspeisung von zeitabhängigen Instrumentenaktivitäten untersucht. Dabei soll keine wie von Cantisani et al. [9] beschriebene Vorarbeit durch den Anwender erfolgen, wodurch die Separationsgeschwindigkeit deutlich höher ist und das Modell in vielen Anwendungsfeldern eingesetzt werden kann. Die Frequenzbereiche der zu trennenden Musikinstrumente können sich häufig überlagern, sodass eine gemeinsame Schätzung aller zu extrahierenden Quellensignale Sinn ergibt, weil sie sich gegenseitig beeinflussen und ihre Separation daher auch von der Zusatzinformation über die anderen Instrumente profitieren kann. Folglich wird hier kein bedingter Ansatz verfolgt, sondern eine Integration der zusätzlichen Instrumentenaktivität ähnlich zum Ansatz von Slizovskaia et al. [126] umgesetzt. Im Gegensatz dazu ist die Zusatzinformation in dieser Arbeit aber zeitabhängig und wird nicht über eine Multiplikation integriert, sondern in zusätzlichen Kanälen angefügt. Dadurch kann das KNN die relevanten Merkmale der Zusatzinformation im Training selbst lernen.

Das in dieser Arbeit untersuchte Separationsmodell basiert auf der Demucs-Architektur [18], die eine UNet-Struktur mit jeweils sechs En- und Decoderblöcken sowie zwei bidirektionalen LSTM-Modulen aufweist. Als Ansatz mit Zeitsignaleinspeisung besitzt Demucs durch die direkte Signalschätzung am Modellausgang eine deutlich höhere, theoretisch erreichbare Separationsqualität als bei der Verwendung von Zeit-Frequenz-Darstellungen bzw. daraus abgeleiteter Masken, die in Abschnitt 5.1.1 beschrieben sind. Auch die tatsächlichen Ergebnisse für die Trennung von Gesang, Bass, Schlagzeug und Rest waren lange Zeit die besten, die in Literaturansätzen für den sehr verbreiteten MUSDB18-

Testdatensatz [108] erzielt wurden. Darüber hinaus ist die Integration von zeitabhängigen Zusatzinformationen in einer Modellarchitektur mit direkter Zeitsignaleinspeisung einfacher, da die Korrelation der zusammengehörigen Zeitabschnitte direkt gegeben ist und die Integration von Zusatzinformationen in sehr feiner Auflösung erfolgen kann.

Zur Schätzung der Signalspuren klassischer Musikinstrumente aus Ensemble-Aufnahmen wird das Demucs-Modell an mehreren Stellen modifiziert. Eine bedeutende Änderung ist die Herabsetzung der Anzahl von En- und Decoderblöcken von sechs auf vier, welche in Vorversuchen zu keiner Verschlechterung der Separationsergebnisse geführt hat, aber die Modellgröße drastisch reduziert. Dies senkt auch den Ressourcenbedarf im Training erheblich, wodurch das Separationsmodell dieser Arbeit mit einer Grafikkarte NVIDIA GeForce RTX 2080 Ti trainierbar ist und kein Array von 16 Grafikkarten benötigt wie Demucs [18]. Zur weiteren Reduzierung der Modellparameter werden anstelle der LSTM-Module die schon in Abschnitt 2.3.1 behandelten bidirektionalen GRU-Module eingesetzt. Sie zeichnen sich durch einen vereinfachten Aufbau mit weniger Parametern gerade für kleinere Datensätze aus [1] und wurden schon erfolgreich in der Musikseparation eingesetzt [80]. Da die Separation hier für Ensembles unterschiedlicher Besetzungen untersucht wird, werden nicht immer vier, sondern allgemein I_{Instr} Musikinstrumente gleichzeitig durch das KNN getrennt. Des Weiteren wird die Einspeisung der zeitabhängigen Instrumentenaktivitäten als Zusatzinformation analysiert. Sie kann theoretisch in jeder Stufe des Separationsmodells eingespeist werden. Um die zusätzliche Information schon in der Datenkompression nutzen zu können, werden in dieser Arbeit ausschließlich die Einspeisepositionen vor den Encoderblöcken oder dem ersten GRU-Modul untersucht. In Abbildung 5.1 ist das Separationsmodell exemplarisch mit der Einspeiseposition vor Encoderblock 4 dargestellt.

Neben der Bezeichnung jedes Blocks ist in Abbildung 5.1 jeweils auch die Anzahl seiner Ausgangskanäle mit angegeben. Fast alle Anzahlen stellen Vielfache der initialen Kanaldimension V_{Sep} des Separationsmodells dar. Pro Encoderstufe wird die Merkmalsanzahl verdoppelt, um viele Eigenschaften der Musikinstrumente einbeziehen zu können. Die En- und Decoderblöcke bestehen jeweils aus mehreren KNN-Bausteinen, deren genaue Zusammensetzung in Abbildung 5.2 illustriert ist. Dabei

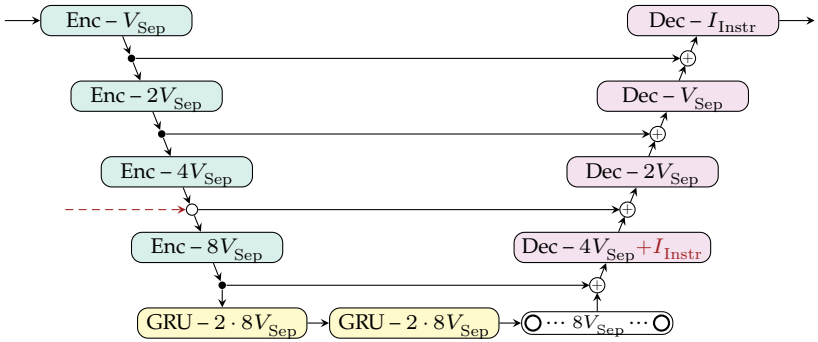
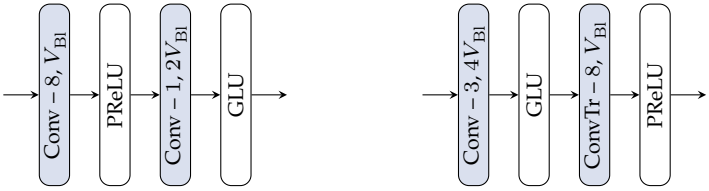


Abbildung 5.1 Schema der Architektur des Separationsmodells aus Encoder- (Enc) und Decoderblöcken (Dec) sowie zwei bidirektionalen GRU-Modulen mit einer nachgeschalteten voll verbundenen Schicht. Die zusätzliche Einspeisung der Instrumentenaktivitäten aller betrachteter Instrumente I_{Instr} ist in rot beispielhaft vor Encoderblock 4 dargestellt.

werden die im Separationsmodell (Abbildung 5.1) angegebenen Anzahlen der Blockausgangskanäle als Dimension V_{Bl} bezeichnet. Im Vergleich zur Demucs-Architektur wird in den En- und Decoderblöcken eine parametrische ReLU-Funktion (PReLU) anstelle der klassischen ReLU-Aktivierung verwendet. Sie berücksichtigt mit

$$\sigma_{\text{PReLU}}(x) = \max(\alpha x, x), \quad \alpha \leq 1 \tag{5.4}$$

auch negative Werte, wobei die Variable α ein im Training trainierbarer Parameter ist. Als zweite Aktivierungsfunktion wird die *Gated Linear Unit* (GLU) [16] eingesetzt. Sie lässt die erste Hälfte der Eingangswerte in Abhängigkeit der anderen Hälfte durch oder sperrt sie, sodass die GLU wie ein bedingter Schalter funktioniert und unwichtige Merkmale ausblenden kann. Pro Encoderblock wird die zeitliche Dimension um den Faktor 4 reduziert, um abstraktere und eine größere Zeitspanne umfassende Merkmale zu extrahieren. Dazu wird kein *Pooling* eingesetzt, sondern jeder Faltungskern mit Kerngröße 8 während der Faltung um die Schrittweite von jeweils 4 Datenpunkten verschoben. Der Decoderblock kann größtenteils als Umkehrung eines Encoderblocks interpretiert werden. Um auf die zeitliche Dimension der darüberliegenden Stufe zu kommen, benötigt jeder Decoder eine transponierte Faltungsschicht, die



(a) Aufbau eines Encoderblocks. **(b)** Aufbau eines Decoderblocks.

Abbildung 5.2 Schema der im Separationsmodell (Abbildung 5.1) verwendeten Blockstrukturen für En- und Decoder, bestehend aus eindimensionalen Faltungsschichten (Conv) bzw. einer transponierten Faltungsschicht (ConvTr) sowie unterschiedlichen Aktivierungsfunktionen. Für alle Faltungsschichten sind die Kerngröße sowie die Anzahl der Ausgangskanäle in Abhängigkeit der Dimension V_{Bl} des Blockausgangs angegeben.

ebenfalls mit einer Schrittweite von 4 arbeitet. Alle anderen Faltungsschichten besitzen standardmäßig eine Schrittweite von 1.

Die Demucs-Architektur arbeitet mit Stereosignalen, wohingegen in dieser Arbeit der allgemeinere Fall von Monosignalen betrachtet wird. Daher liegt das betrachtete Zeitsignal am Eingang des Separationsmodells als ein Vektor an. Aufgrund der sehr großen Datenmengen, die bei direkter Verwendung der Zeitsignale eingespeist werden müssen, steigt der Speicherbedarf im Training enorm. Durch die Einspeisung von kleinen Signalsegmenten kann dieser deutlich verringert werden, weshalb die Eingangsdimension des Separationsmodells auf 65 536 Datenpunkte gesetzt wird. Diese Eingangslänge umfasst bei einer Abtastfrequenz von 48 kHz ungefähr 1,37 s, was deutlich unter den bei Demucs analysierten 10 s liegt. Vorversuche mit der achtfachen Eingangslänge von etwa 10,92 s führten bei dem hier vorgestellten KNN aber zu keiner signifikanten Verbesserung der Separationsergebnisse. Folglich werden die zu trennenden Zeitsignale vor der Einspeisung in Segmente von 65 536 Datenpunkten aufgeteilt, um den Ressourcenbedarf im Training klein zu halten. Wie bei Demucs führt eine *Batch*-Normalisierung zu sehr schlechten Separationsergebnissen, da Rauschen stark verstärkt auftritt. Als alternative Regularisierung wird hier *Dropout* mit einer Rate von 0,5 in der voll verbundenen Schicht nach den GRU-Modulen verwendet. Insgesamt besitzt das Separationsmodell etwa 41 Millionen Parameter, je nach I_{Instr} ein paar tausend Parameter mehr oder weniger.

Auch mit der zusätzlichen Einspeisung der zeitabhängigen Instrumentenaktivität bleibt die Modellgröße beinahe identisch, da im Gegensatz zu Literaturansätzen mit Kontrollnetzen und FiLM oder LaSAFT keine zusätzlichen Schichten zur Integration der Zusatzinformation benötigt werden. Die zusätzliche Information wird, wie oben bereits kurz beschrieben und in Abbildung 5.1 mit der roten gestrichelten Linie illustriert, vor einem Encoderblock an die Eingangsdaten angefügt, indem die Instrumentenaktivitäten zusätzliche Kanäle bilden. Dadurch kann das KNN die für die Separation wichtigen Informationen extrahieren und in seinen Merkmalen miteinbeziehen, aber die Merkmale des zu trennenden Musiksignals werden nicht automatisch durch Multiplikation oder Addition der Zusatzinformation manipuliert, sodass keine Verzerrung der Quellensignale auftritt. Wegen der stufenweisen Reduktion der Zeitauflösung im Encoder müssen die zeitabhängigen Instrumentenaktivitäten ebenfalls komprimiert werden, sofern sie nicht direkt vor Encoderblock 1 eingespeist werden. Je nach Einspeisepunkt wird die Zeitauflösung der Zusatzinformation um den Faktor $4^{j_{\mathrm{Bl}}}$ reduziert, damit sie zum Ausgang des vorhergehenden Encoderblocks j_{Bl} passt. In Abbildung 5.1 liegt der Faktor beispielsweise bei 64, sodass noch 1024 Zeitwerte pro Instrument als Zusatzinformation eingespeist werden. Die Anzahl der zusätzlich eingespeisten Kanäle beträgt I_{Instr}. Falls die Einspeisung nicht vor Encoderblock 1 erfolgt, übermittelt die jeweilige *Skip*-Verbindung auch die Zusatzinformation an die Addition vor dem Decoderblock j_{Bl}. Damit in dieser Addition keine Dimensionskonflikte auftreten, muss der Ausgang des vorhergehenden Decoderblocks um I_{Instr} Merkmale erhöht werden, was ebenfalls in Abbildung 5.1 dargestellt ist.

Im Rahmen dieser Arbeit werden sowohl die bisher beschriebenen Gesamtmodelle zur gemeinsamen Schätzung aller Quellensignale als auch Einzelmodelle untersucht, die das separierte Signal jeweils eines Musikinstruments extrahieren. Solche unabhängigen Einzelmodelle ermöglichen eine flexible Zusammenstellung der zu trennenden Musikinstrumente und sind einfach erweiterbar. Der Nachteil von unabhängigen KNN ist die sehr lange Trainingsdauer, da jedes Modell nacheinander mit dem gleichen Datensatz trainiert werden muss. Darüber hinaus können die Einzelmodelle nicht von Zusatzinformationen anderer Quellen profitieren. Für das gemeinsame Separationsmodell werden Fälle mit 13 und 3 zu

trennenden Musikinstrumenten analysiert. Im Falle der unabhängigen Einzelmodelle ist der Parameter I_{Instr} dagegen stets 1. Deshalb müssen in den Einzelmodellen weniger Merkmale extrahiert werden, sodass die Variable V_{Sep} von 100 im Gesamtmodell auf 50 für jedes unabhängige Separationsmodell halbiert wird. Folglich reduziert sich die Modellgröße von ungefähr 41 Millionen Parametern im Gesamtmodell auf etwa 11,5 Millionen Parameter für ein Einzelmodell. Bei vier oder mehr zu separierenden Quellen benötigt das System aus unabhängigen Einzelmodellen somit mehr Speicherplatz. Die Rechenzeit liegt bei etwa 65 ms für das Gesamtmodell, wohingegen jedes Einzelmodell ungefähr 40 ms für die Schätzung des extrahierten Quellensignals benötigt.

Alle Separationsmodelle dieser Arbeit sind in Tensorflow implementiert und werden mithilfe des Adam-Optimierers [66] sowie der quadratischen Kostenfunktion MSE trainiert. Der Adam-Optimierer ist dem SGD besonders beim Lernen von seltenen Merkmalen überlegen [66], sodass er für den hier betrachteten Fall von Ensembles unterschiedlicher Besetzungen mit teilweise selten vorkommenden Instrumenten geeignet ist. Das Training mit einer Lernrate von $3 \cdot 10^{-4}$ wird in *Batches* von 32 Musiksignalsegmenten über maximal 500 Epochen durchgeführt. Um unnötige Trainingszeit und eine Überanpassung des KNN zu verhindern, kommt das sogenannte *Early Stopping* zum Einsatz, bei dem das Training nach 50 Epochen ohne Verbesserung der durchschnittlichen Separationsergebnisse für die im Training unbekannten Validierungsdaten abgebrochen wird. Nach Beendigung des Trainings wird immer das Modell mit dem besten Separationsergebnis für die Validierungsdaten verwendet. Für alle Zufallsfunktionen des Separationsmodells und des Trainings wird der gleiche Startpunkt (*Seed*) von 0 definiert, damit die verschiedenen Modellvarianten besser vergleichbar sind und ihre Separationsqualität nicht von unterschiedlichen Initialisierungen verfälscht wird.

5.3 Datensatz

Für die hier untersuchte, datenbasierte Separation polyphoner Musiksignale ist ein ausreichend großer Datensatz notwendig, der Einzelspuren unterschiedlicher Musikinstrumente aus Ensemble-Aufnahmen enthält. Viele Datensätze der Literatur beinhalten nur eine begrenzte Menge an

Komponenten, so auch der bereits angeführte, weit verbreitete MUSDB18-Datensatz [108] mit den vier Komponenten Gesang, Schlagzeug, Bass und Rest. Der in Kapitel 4 verwendete MusicNet-Datensatz [137] enthält zwar verschiedene Musikinstrumente, aber keine getrennten Quellensignale, sodass er für das überwachte Training der Separation ebenfalls ungeeignet ist. Ein ausreichend großer Datensatz, der die geforderten Eigenschaften aufweist, ist der *University of Rochester Musical Performance* (URMP) Datensatz [75]. Er umfasst 44 klassische Musikstücke, die von kleinen Ensembles unterschiedlicher Besetzungen eingespielt wurden. Die 11 Duos, 12 Trios, 14 Quartette und 7 Quintette setzen sich aus den insgesamt 13 betrachteten Musikinstrumenten zusammen. Dabei sind sowohl Streich- als auch Holz- und Blechblasinstrumente enthalten, um verschiedene Instrumentenklassen abzubilden. Die 44 Musikstücke haben eine Gesamtlänge von rund 80 min und sind mit der Abtastfrequenz 48 kHz aufgenommen. Darüber hinaus sind zu jedem Stück die Noten der aktiven Instrumente mit Anschlagszeitpunkt, Notenbezeichnung und Notenlänge bekannt.

In den Besetzungen der URMP-Ensembles kommen teilweise zwei oder mehr Instrumente des gleichen Instrumententyps vor. Diese können vom Separationsmodell dieser Arbeit nicht getrennt werden, weil das KNN das monaurale Eingangssignal auf Basis der Instrumentencharakteristika separiert. Deshalb werden die Einzelspuren des Instrumententyps für das jeweils betroffene Musikstück addiert und das resultierende Signal als neues Quellensignal übernommen. Das hat zur Folge, dass für manche Instrumente, die eigentlich immer nur einen Ton gleichzeitig spielen können, auch mehrstimmige Quellensignale auftreten. Folglich liegt der Fokus des überwachten Trainings automatisch mehr auf dem Instrumentenklang, was positiv für die gewünschte Separationsaufgabe ist. Nach Durchführung aller notwendigen Zusammenlegungen enthält der modifizierte URMP-Datensatz 4 Solos, 12 Duos, 20 Trios und 8 Quartette. Diese Musikaufnahmen werden in 36 Trainings-, 5 Validierungs- (URMP-Nummer 5, 12, 17, 24 und 40) und 3 Teststücke (URMP-Nummer 8, 18 und 41) aufgeteilt. Alle Teildatensätze enthalten dabei mindestens eine Duo-, eine Trio- und eine Quartettaufnahme, um die Varianz der Ensemble-Besetzungen abzudecken.

Verglichen mit Datensätzen für überwachte Trainingsprozesse anderer Domänen, wie z. B. in der Bildverarbeitung, ist der URMP-Daten-

satz relativ klein. Deshalb wird die im Training verfügbare Datenmenge durch eine Augmentierung des Datensatzes erweitert. Es gibt zahlreiche Augmentierungsstrategien mit unterschiedlich starker Veränderung der ursprünglichen Daten des Datensatzes. In dieser Arbeit ist aufgrund der vielen isolierten Instrumentensignale eine vergleichsweise einfache Augmentierung möglich, indem zufällige Einzelspuren unterschiedlicher Instrumente und Musikstücke zu neuen „Aufnahmen“ addiert werden. Sie stellen für das menschliche Ohr keine schön klingende Musik dar, da die Einzelspuren aus unterschiedlichen Musikstücken stammen und daher oft harmonisch und rhythmisch nicht zueinander passen, aber die Separation verschiedener Instrumententypen kann auch mithilfe solcher inkohärenten Aufnahmen trainiert werden. Theoretisch sind durch diese Augmentierungsstrategie sehr viele inkohärente Musikstücke generierbar, sodass die Trainingsdatenmenge erheblich vergrößert werden könnte. Um auch harmonische Zusammenhänge zwischen zusammen spielenden Instrumenten in der Separation berücksichtigen zu können, sollten aber ausreichend viele originale Musikaufnahmen im Datensatz enthalten sein. Deshalb wird die Trainingsdatenmenge durch die Augmentierung in dieser Arbeit nur ungefähr verdoppelt.

Eine Alternative für mehr kohärente Musikaufnahmen ist die synthetische Generierung von Musiksignalen [86, 93], durch welche eine große Zahl von zusätzlichen Spuren und Musikstücken erzeugt werden kann. Wie schon in Abschnitt 5.1.2 ausgeführt, ist die Generalisierung bei synthetischen Musiksignalen aber sehr gering, da die synthetischen Töne bei identischen Parametern immer exakt gleich klingen und sie somit keine Variabilität aufweisen. Folglich sind sie für lernende Verfahren oft ungeeignet, da das Risiko einer Überanpassung an diese spezifischen Instrumentencharakteristiken groß ist. Falls abstraktere Informationen wie die vorkommenden Noten extrahiert werden sollen, können synthetische Daten im Training eingesetzt werden [A5], da die relevanten Merkmale in den synthetischen und den realen Daten ähnlich sind. Im Falle der Separation kommt es aber auf den genauen Signalverlauf an, wodurch sich eine Optimierung auf synthetische Signale im Training negativ auf die Separationsqualität von realen Aufnahmen auswirkt. Daher wird die Generierung von synthetischen Signalen hier nicht berücksichtigt.

Die Augmentierung durch Addition von zufälligen Einzelspuren kann für alle 13 Musikinstrumente des URMP-Datensatzes durchgeführt werden. Soll dagegen hauptsächlich eine spezielle Ensemble-Besetzung separiert werden, macht eine gezielte Augmentierung Sinn, die nur Einzelspuren von Instrumenten des gewünschten Ensembles neu mischt. Im Rahmen dieser Arbeit werden sowohl Separationsmodelle für alle 13 Musikinstrumente als auch reduzierte Modelle für ein Beispieltrio aus Violine, Trompete und Flöte untersucht. Dieses Trio entspricht der Besetzung des Teststücks Nummer 18 und enthält mit je einem Streich-, Holzblas- und Blechblasinstrument alle Instrumentenklassen des URMP-Datensatzes. Um beide Modelltypen gut miteinander vergleichen zu können, werden die identischen augmentierten Trainingsdaten verwendet, welche ausschließlich durch die Addition von Einzelspuren der drei Trioinstrumente erzeugt werden. Vor ihrer Addition werden die zufällig ausgewählten Instrumentenspuren zeitlich um einen zufälligen Wert verschoben und mit einem zufälligen Verstärkungsfaktor zwischen 0,7 und 1,3 gewichtet, damit eine möglichst hohe Varianz im augmentierten Datensatz enthalten ist. Dadurch treten die schon im originalen URMP-Datensatz enthaltenen Instrumentenspuren in veränderter Form auf, was die Gefahr von Überanpassung reduziert. Darüber hinaus kommt keine Kombination der Einzelspuren doppelt vor. Mit diesem Vorgehen werden vor Beginn der Experimente 50 zusätzliche Musikstücke des definierten Beispieltrios generiert, die in dieser Arbeit aus Gründen der besseren Vergleichbarkeit in allen Trainingsprozessen mit augmentierten Daten zum Einsatz kommen. Sie werden in 45 Trainings- und 5 Validierungsstücke aufgeteilt, sodass der augmentierte Datensatz 81 Trainings-, 10 Validierungs- und 3 Teststücke enthält. Vorversuche zeigen, dass die Separationergebnisse durch das Training mit dem augmentierten Datensatz nicht nur für die Instrumente der Augmentierung, sondern für fast alle betrachteten Instrumente verbessert werden. Deshalb wird in dieser Arbeit stets der augmentierte Datensatz verwendet.

Jedes Musikstück wird, wie in Abschnitt 5.2 beschrieben, in Signalsegmente mit 65 536 Datenpunkten aufgeteilt, die in das Separationsmodell eingespeist werden. Während des Trainings erfolgt diese Segmentierung ausgehend von einer zufälligen zeitlichen Verschiebung zwischen 0 und 65 536 Datenpunkten, die pro Musikstück und Epoche neu generiert

wird. Mit dieser Verschiebung wird das erste Signalsegment jedes Musikstücks extrahiert. Alle weiteren Segmente werden daran anschließend, aber ohne Überlappung zu den anderen Segmenten extrahiert. Durch die zeitliche Variation der Segmente wird die Gefahr von Überanpassung des Separationsmodells an die Trainingsdaten reduziert. Für den URMP-Datensatz entsprechen 65 536 Werte einer Zeitdauer von ungefähr 1,37 s. Diese ins Modell eingespeiste Zeitspanne kann durch Unterabtastung der Musiksignale vergrößert werden. Aufgrund des damit verbundenen Informationsverlustes und keinen signifikant besseren Separationsergebnissen in einzelnen Vorversuchen wird in dieser Arbeit auf die Analyse einer niedrigeren Abtastrate verzichtet.

Die zeitabhängige Instrumentenaktivität, welche als Zusatzinformation für die Separation dient, wird für jedes der berücksichtigten I_{Instr} Musikinstrumente aus den im URMP-Datensatz angegebenen Notenverläufen entnommen. Daraus ergibt sich pro Musikstück für jedes Instrument ein binäres Aktivitätssignal mit der Abtastrate 48 kHz, das in allen Zeitpunkten eine 1 aufweist, in denen das entsprechende Instrument eine Note spielt. Um als Vektor in das Separationsmodell integriert werden zu können, muss das Signal wie in Abschnitt 5.2 beschrieben auf die zum Einspeiseort passende zeitliche Dimension reduziert werden. Auch wenn die im URMP-Datensatz angegebenen Noteninformationen Fehler bzw. kleine zeitliche Ungenauigkeiten enthalten können, werden sie als korrekte Zusatzinformation interpretiert und in jedem Training der Separationsmodelle verwendet. Für die reale Anwendung der Separation ist die Instrumentenaktivität jedoch nicht bekannt, weshalb ein Detektionssystem wie z. B. der in Kapitel 4 vorgestellte Ansatz vorgeschaltet werden muss, um die Zusatzinformation zu schätzen. Diese Schätzung ist in der Regel fehlerhaft, weshalb die Robustheit der Separationsansätze in Abschnitt 5.5 auf Basis von simulativ erzeugten Fehlern untersucht wird. Die Fehler werden durch Invertierung eines definierten Prozentsatzes der fehlerfreien Instrumentenaktivitätswerte realisiert. Dabei werden die genauen Zeitpunkte der Invertierung zufällig gewählt und alle Aktivitätssignale der Instrumente separat behandelt.

5.4 Multi-Task-Ansatz zur Separation

Zunächst werden Separationsmodelle ohne zusätzliche Einspeisung der Instrumentenaktivität untersucht. Neben den in Abschnitt 5.2 beschriebenen Ansätzen eines Gesamtmodells für alle zu trennenden Musikinstrumente und einem System aus Einzelmodellen wird in diesem Abschnitt mit dem Multi-Task-Ansatz eine dritte Modellarchitektur analysiert. Denn Multi-Task-Ansätze der Literatur [52, 84] erzielen, wie in Abschnitt 5.1.2 beschrieben, bessere Ergebnisse für zwei fusionierte MIR-Aufgaben. Da die Extraktion jeder zu separierenden Instrumentenspur aufgrund der instrumentenspezifischen Charakteristiken auch als eigenständige Aufgabe interpretiert werden kann, repräsentiert die Separation mehrerer Spuren ebenfalls ein Multi-Task-Problem.

Die Architektur des hier untersuchten Multi-Task-Ansatzes unterscheidet sich nur im obersten Decoderblock von der Struktur des Gesamtmodells ohne Zusatzinformation in Abbildung 5.1. Statt des einen Ausgangsblocks mit I_{Instr} Ausgangskanälen wird die Extraktion jedes Instruments mithilfe eines separaten Decoderblocks mit einem Ausgangskanal umgesetzt. Durch die parallele Anordnung der insgesamt I_{Instr} Decoderblöcke in der obersten Decoderschicht ist die Ausgangsdimension der getrennten Quellensignale identisch zu der des Gesamtmodells. Das KNN des Multi-Task-Ansatzes besitzt nur unwesentlich mehr Parameter als das Gesamtmodell und ist damit deutlich kleiner als ein Separationssystem von vielen Einzelmodellen. Aufgrund der instrumentenspezifischen Decoderblöcke der letzten Schicht kann das Multi-Task-Modell stärker an die betrachteten Musikinstrumente angepasst werden. Dabei berücksichtigt es trotzdem die Zusammenhänge zwischen den Instrumenten, da die in der sonstigen Modellstruktur extrahierten Merkmale für alle Instrumente gleich sind. Folglich stellt der Multi-Task-Ansatz eine Mischung des Gesamtmodells und des Systems aus Einzelmodellen dar.

Im Training werden die Parameter des Multi-Task-Modells mithilfe der angepassten Gütefunktion

$$J_{\mathrm{MT}} = \sum_{i=1}^{I_{\mathrm{Instr}}} \chi_i \, J_i + J_{\mathrm{reg},i} \tag{5.5}$$

optimiert, welche die mit den Faktoren χ_i gewichtete Summe der Gütefunktionen J_i jedes Instruments i beinhaltet. Analog zu den anderen Ansätzen wird hier für jede Gütefunktion J_i der MSE verwendet. Darüber hinaus enthält die angepasste Multi-Task-Gütefunktion einen Regularisierungsterm $J_{\mathrm{reg},i}$, um ungewollte Effekte wie Signalrauschen oder bestimmte Modellgewichte nicht zu groß werden zu lassen. Der Regularisierungsterm kann dabei abhängig oder unabhängig vom Instrument i sein. Ein sehr einfacher Fall der Multi-Task-Gütefunktion J_{MT} resultiert aus der statischen Definition von $\chi_i = 1$ und $J_{\mathrm{reg},i} = 0$. Dadurch werden die Gütefunktionen aller zu trennenden Instrumente in der Optimierung gleich gewichtet und keine zusätzliche Regularisierung verfolgt, sodass die Gesamtfunktion einer einfachen Summe entspricht. Der Vorteil des Multi-Task-Ansatzes liegt aber unter anderem in der angepassten Gewichtung der einzelnen, parallel gelernten Aufgaben. Diese wird beispielsweise durch die *Multi-Task Uncertainty* [60] erreicht, welche die Parameter durch

$$\chi_i = \frac{1}{2\sigma_i^2} \qquad J_{\mathrm{reg},i} = \ln\left(\sigma_i\right) = -0{,}5 \ln\left(\chi_i\right) \tag{5.6}$$

definiert. Bei diesem Ansatz hängen die Gewichtungen χ_i und die instrumentenabhängige Regularisierung $J_{\mathrm{reg},i}$ von der Varianz σ_i der als Gauß-Verteilung angenommenen Wahrscheinlichkeitsverteilung des Ausgangs ab. Die resultierende Gütefunktion J_{MT} kann entweder in Abhängigkeit der Gewichtungen χ_i oder der Varianzen σ_i formuliert werden, wobei die entsprechenden Variablen dann trainierbare Parameter repräsentieren.

Alle Separationsmodelle werden ohne Zusatzeinspeisung der Instrumentenaktivität mithilfe des in Abschnitt 5.3 vorgestellten augmentierten URMP-Datensatzes trainiert. Da die Augmentierung gezielt für die Besetzung Violine, Trompete und Flöte vorgenommen wurde, treten die Unterschiede der Separationsergebnisse für Modelle dieses Beispieltrios am stärksten hervor. Deshalb werden im vorliegenden Abschnitt ausschließlich diese Trio-Modelle analysiert. Im Training des Multi-Task-Ansatzes werden zur Optimierung einmal die einfache Gütefunktion ($\chi_i = 1$ und $J_{\mathrm{reg},i} = 0$) und einmal die auf der *Multi-Task Uncertainty* basierende Formel angewendet. Die Separationsergebnisse dieser beiden Varianten sind in Tabelle 5.1 den Ergebnissen für das Gesamtmodell des Beispieltrios und denen des Separationssystems aus den jeweiligen

Tabelle 5.1 SI-SDR (in dB) unterschiedlicher Separationsmodelle mit und ohne Multi-Task- (MT-) Architektur für die Trennung der drei Instrumente Violine, Trompete und Flöte im URMP-Testdatensatz.

	Flöte	Tromp.	Violine	∅
MT mit $\chi_i = 1$ und $J_{\text{reg},i} = 0$	0,32	0,86	–1,82	–0,21
MT mit *MT Uncertainty*	0,26	0,99	–2,59	–0,45
Gesamtmodell ohne MT	**2,60**	–0,26	**–0,07**	0,76
System aus Einzelmodellen	1,99	**3,38**	–1,00	**1,45**

Einzelmodellen gegenübergestellt. Jedes Ergebnis entspricht dabei dem Mittelwert der SI-SDR-Werte aller Stücke des URMP-Testdatensatzes.

Aus Tabelle 5.1 wird deutlich, dass die untersuchten Multi-Task-Ansätze Musiksignale insgesamt schlechter separieren als die in Abschnitt 5.2 vorgestellten Modelle. Vor allem die Flöten- und Violinensignale sind für das niedrigere durchschnittliche SI-SDR der Multi-Task-Ansätze verantwortlich. Der Vorteil eines ausgewogeneren Trainings durch die trainierbaren Gewichte kann in diesem Fall nur teilweise bestätigt werden, weil das Ergebnis für Flöte und Violine zwar schlechter, aber immerhin für die Trompete besser als im Falle des Gesamtmodells ist. Durch die separaten Decoderblöcke wird die Trennung der Trompete in der Multi-Task-Architektur gestärkt und in der Optimierung weniger von der Separation der anderen beiden Instrumente überlagert. Noch stärker tritt dieser Effekt im Falle der Einzelmodelle hervor, die wesentlich höhere SI-SDR-Werte für die Trompete erzielen als alle anderen Ansätze.

Auch wenn das SI-SDR zur Bewertung der getrennten Instrumentensignale in den meisten Fällen ausreicht, ermöglicht die Analyse der Metriken SI-SIR und SI-SAR eine detailliertere Einschätzung der Separationsqualität und ihrer Ursachen. Deshalb sind die Durchschnittswerte der drei Metriken für alle untersuchten Separationsansätze in Abbildung 5.3 dargestellt. In allen betrachteten Ansätzen ist die Trennung der Instrumentenstimmen erfolgreich, was an den hohen SI-SIR-Werten abgelesen werden kann. Die extrahierten Signale enthalten jedoch viele Artefakte, wodurch das SI-SAR und damit auch das SI-SDR für alle Ansätze sehr niedrig ist. Diese Artefakte resultieren unter anderem aus

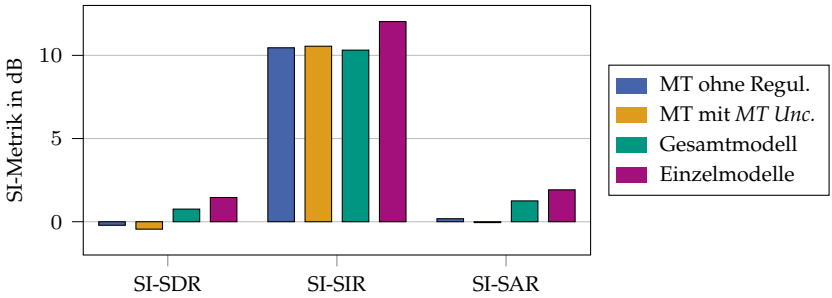


Abbildung 5.3 Vergleich der durchschnittlichen Separationsergebnisse mit und ohne Multi-Task- (MT-) Architektur für die Trennung der drei Instrumente Violine, Trompete und Flöte im URMP-Testdatensatz.

der eingesetzten Netzarchitektur mit direkter Zeitsignalschätzung, da schon kleine Abweichungen von den Originalwerten der Quellensignale als Artefakte aufgefasst werden. Im Vergleich der untersuchten Modelle bestätigt sich auch bei Analyse der beiden Metriken SI-SIR und SI-SAR die Unterlegenheit der Multi-Task-Ansätze, weshalb sie im Folgenden nicht weiter betrachtet werden.

5.5 Separationsergebnisse mit simulierter Zusatzinformation

Die Separation monauraler Ensemble-Aufnahmen trennt die Instrumentensignale auf Basis ihrer charakteristischen Klangeigenschaften. Um die Separation durch zusätzliche Information zu unterstützen, werden die zeitabhängigen Instrumentenaktivitäten aller zu trennenden Quellen als weitere Eingangssignale in das Separationsmodell eingespeist. Die Auswirkungen der zusätzlichen Einspeisung in verschiedenen Netzebenen werden zunächst in Abschnitt 5.5.1 für die Architektur eines gemeinsamen Gesamtmodells untersucht. Anschließend werden in Abschnitt 5.5.2 die Auswirkungen für ein System aus mehreren Einzelmodellen mit Zusatzeinspeisung analysiert. In beiden Fällen erfolgt die Analyse mithilfe der aus dem Datensatz bekannten Instrumentenaktivitäten, die direkt oder mit simulierten zufälligen Fehlern eingespeist werden.

5.5.1 Gemeinsames Separationsmodell

Das in Abschnitt 5.2 vorgestellte gemeinsame Separationsmodell schätzt die Quellensignale aller I_{Instr} zu trennenden Musikinstrumente gleichzeitig und profitiert daher von den extrahierten Merkmalen und Zusatzinformationen der anderen Instrumente. Auch wenn die Quellenanzahl I_{Instr} beliebig wählbar ist, muss sie vorab definiert werden. Das KNN schätzt dann immer I_{Instr} Signale der zugehörigen Quellen. Aufgrund der 13 Instrumente im verwendeten URMP-Datensatz wird als erstes ein Separationsmodell mit $I_{\text{Instr}} = 13$ getrennten Ausgangssignalen betrachtet. Es wird mit dem in Abschnitt 5.3 beschriebenen, augmentierten URMP-Datensatz und den in Abschnitt 5.2 definierten Parametern und Implementierungsdetails trainiert.

Die zusätzliche Integration der zeitabhängigen Instrumentenaktivität wird an allen möglichen Einspeisepunkten des Encoders untersucht, wobei der Einspeiseort jeweils nach dem darauffolgenden Encoderblock benannt ist. Nach erfolgreichem Training wird die Separation aller Modellvarianten mithilfe des in Abschnitt 5.3 definierten Testdatensatzes bewertet. Da die drei Musikstücke des Testdatensatzes nur sechs Musikinstrumente enthalten, sind im Folgenden immer nur die Separationsergebnisse dieser sechs Instrumente angegeben und diskutiert. Alle getrennten Signale der anderen sieben Instrumente enthalten fast nur Stille, weshalb sie nicht näher ausgewertet werden. Die Ergebnisse der gemeinsamen Separationsmodelle mit unterschiedlichen Einspeisepositionen sind in Tabelle 5.2 anhand der instrumentenspezifischen SI-SDR-Werte aufgelistet. Darüber hinaus sind die Ergebnisse des Modells ohne Zusatzeinspeisung (-) angegeben. Der Fall ohne Separation (‚Mix'), bei dem jedes Ausgangssignal dem gemischten Eingangssignal entspricht, repräsentiert die schlechteste Separation. Als weitere Referenz zur Integration der zeitabhängigen Aktivitätsinformation ist in Tabelle 5.2 die Multiplikation der ohne Zusatzinformation geschätzten Instrumentensignale mit der jeweiligen Instrumentenaktivität angegeben (‚Mult.'). Durch die Multiplikation werden alle Signalanteile unterdrückt, die außerhalb der entsprechenden Instrumentenaktivität liegen.

Aus den durchschnittlichen SI-SDR-Werten der Modellvarianten in Tabelle 5.2 geht klar hervor, dass die Separation der Musiksignale durch die Einspeisung der zeitabhängigen Instrumentenaktivität verbessert

Tabelle 5.2 SI-SDR (in dB) aller Instrumente des URMP-Testdatensatzes für gemeinsame Separationsmodelle ohne (-) und mit Zusatzeinspeisung an unterschiedlichen Einspeiseorten. Als Referenzen dienen das nicht separierte Eingangssignal (‚Mix') und die Multiplikation der zeitabhängigen Instrumentenaktivität mit der Schätzung ohne sie (‚Mult.').

	Fagott	Flöte	Oboe	Sax.	Tromp.	Violine	∅
-	–7,60	3,06	–0,99	–6,62	–0,28	–1,07	–2,25
Enc. 1	**1,31**	1,89	–0,75	–5,86	**3,95**	**1,62**	0,36
Enc. 2	0,26	0,82	–2,28	–3,80	3,80	–1,51	–0,45
Enc. 3	–2,77	3,47	–3,04	–3,80	1,60	–1,68	–1,04
Enc. 4	0,86	**4,74**	**–0,42**	**–3,53**	1,36	0,65	**0,61**
GRU	0,62	4,01	–1,38	–3,80	1,00	–0,16	0,05
Mix	–3,51	1,29	–11,29	–6,17	–4,12	–4,98	–4,80
Mult.	–7,60	3,17	–0,96	–6,59	–0,18	–1,12	–2,22

wird. Die erzielten Ergebnisse steigen dabei für alle Einspeisepositionen um mindestens 1,21 dB gegenüber dem Fall ohne Zusatzinformation. Darüber hinaus übertreffen die Ergebnisse mit zusätzlicher Einspeisung die SI-SDR-Werte für das originale Eingangssignal in allen betrachteten Instrumenten, was bei der Schätzung ohne Zusatzinformation nicht für das Fagott und das Saxofon gilt.

Zwischen den Einspeiseorten sind deutliche Unterschiede zu verzeichnen. Die Einspeisung vor Encoderblock 2 und 3 ist beispielsweise für die Instrumente Fagott, Oboe und Violine nicht zufriedenstellend, weshalb sie auch die im Durchschnitt schlechtesten SI-SDR-Werte besitzen. Bei der Integration der Zusatzinformation vor Encoderblock 4 werden dagegen mit einem durchschnittlichen SI-SDR von 0,61 dB die besten Separationsergebnisse erzielt. Daraus lässt sich ableiten, dass die zeitabhängige Instrumentenaktivität als Zusatzinformation am wertvollsten ist, wenn sie mit komplexen, abstrakten Merkmalen kombiniert wird, die in den tiefen Stufen der UNet-Struktur und der latenten Schicht vorkommen. Sie umfassen aufgrund der Datenkompression automatisch auch immer eine größere Zeitspanne als Merkmale in den ersten Stufen. Um diese abstrakten Merkmale noch vor dem ersten GRU-Modul mit der Zusatzinformation verknüpfen zu können, eignet sich die Einspeisung

vor dem letzten Encoderblock am besten, sodass die Verknüpfung im letzten Encoderblock realisiert wird. Folglich ist die Separationsqualität bei Einspeisung vor Encoderblock 4 auch etwas höher als vor dem ersten GRU-Modul, dessen Merkmale noch komprimierter sind. Durch die Reduktion der Zeitauflösung pro Stufe enthält die vor dem Encoderblock 4 eingespeiste Zusatzinformation noch 1024 Zeitwerte, was einer Zeitauflösung von 1,33 ms entspricht.

Wird die zeitabhängige Instrumentenaktivität erst nach der Schätzung aller Instrumentensignale durch eine Multiplikation integriert, kann sie die ohne Zusatzinformation separierten Quellensignale nur noch marginal verbessern. Im Falle der Violine verschlechtert sich das SI-SDR sogar leicht, weil durch die harte Signalunterdrückung der Multiplikation Artefakte auftreten. Insgesamt lässt sich aus der nur sehr geringen Verbesserung durch die Multiplikation entnehmen, dass die vom KNN geschätzten Instrumentensignale in den nicht aktiven Passagen kaum Anteile anderer Instrumente enthalten, die von der Multiplikation unterdrückt werden könnten. Das ist eine sehr positive Eigenschaft des vorgestellten Separationsansatzes. Die aktiven Passagen jedes Instruments werden von der Multiplikation nicht beeinflusst, sodass die Separationsergebnisse nicht wesentlich verbessert werden können.

Im Vergleich zur Einspeisung von statischen Zusatzinformationen über die im Musikstück vorhandenen Instrumente, die in den meisten Literaturansätzen wie z. B. von Slizovskaia et al. [126] umgesetzt wird, verbessert die in dieser Arbeit vorgeschlagene Einspeisung der zeitabhängigen Instrumentenaktivität die Separation. Exemplarisch sind die SI-SDR-Werte der statischen und zeitabhängigen Einspeisung vor Encoderblock 4 in Abbildung 5.4 gegenübergestellt. Dabei wird in beiden Fällen das gleiche, mit zeitabhängiger Zusatzinformation trainierte Separationsmodell verwendet. Im Falle der statischen Information enthält der eingespeiste Aktivitätsvektor eines Instruments nur Einsen, wenn es im Musikstück spielt, und nur Nullen, wenn es im Stück nicht vorkommt. Damit wird ein statischer Wert pro Instrument imitiert. Anhand der im Durchschnitt um 0,53 dB besseren Separationsqualität des zeitabhängigen Falles wird deutlich, dass die Zeitabhängigkeit der Instrumentenaktivität wichtig ist und im Modell für eine gezieltere Separation sorgt. Dies zeigt sich auch in den Verbesserungen der meisten Instrumentensignale,

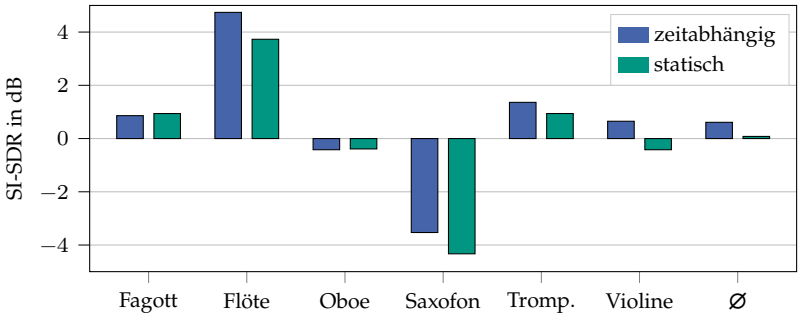


Abbildung 5.4 Vergleich der Separationsergebnisse des URMP-Testdatensatzes bei zeitabhängiger und statischer Information über aktive Instrumente, die im gemeinsamen Separationsmodell vor Encoderblock 4 eingespeist wird.

mit Ausnahme von Fagott und Oboe. Für alle anderen Einspeiseorte ist die durchschnittliche Verbesserung der Separation durch die Zeitabhängigkeit sogar noch höher. Die genauen Zahlenwerte mit zeitabhängiger und statischer Zusatzinformation aller Einspeisepositionen sind in Anhang B aufgelistet. Dort sind für alle bisher diskutierten, gemeinsamen Separationsmodelle sowohl das SI-SDR in Tabelle B.1 als auch das SI-SIR in Tabelle B.2 und das SI-SAR in Tabelle B.3 angegeben.

Falls die zu trennende Ensemble-Besetzung bekannt ist und nur wenige Musikinstrumente umfasst, kann das KNN durch Anpassung der Anzahl aller zu trennenden Quellen I_{Instr} reduziert werden. Dadurch wird das Separationsmodell gezielter trainiert und die Modellparameter hinsichtlich der definierten Instrumente optimiert, sodass eine Verbesserung der Separationsqualität zu erwarten ist. Dieser Fall der auf eine kleine Besetzung angepassten Separation wird mithilfe des schon im Datensatz verwendeten Beispieltrios aus Violine, Trompete und Flöte untersucht. Das resultierende KNN wird mit dem identischen augmentierten URMP-Datensatz trainiert wie das zuvor analysierte Modell aller 13 Instrumente. Die zugehörigen Separationsergebnisse für unterschiedliche Einspeisepositionen der zeitabhängigen Instrumentenaktivität sind in Tabelle 5.3 durch die SI-SDR-Werte des Testdatensatzes angegeben. Als Referenz dient wie beim Gesamtmodell mit 13 Instrumenten die Multiplikation der Zusatzinformation mit den ohne sie geschätzten In-

Tabelle 5.3 SI-SDR (in dB) der Instrumente des definierten Beispieltrios für gemeinsame Separationsmodelle dieses Trios ohne (-) und mit Zusatzeinspeisung an unterschiedlichen Orten. Als Referenzen dienen die Multiplikation der Zusatzinformation am Ausgang und die Einspeisung statischer Instrumentenaktivität vor Encoderblock 4.

	Flöte	Tromp.	Violine	∅
-	2,60	–0,26	–0,07	0,76
Encoderbl. 1	**5,06**	0,77	1,69	2,51
Encoderbl. 2	4,00	1,20	1,40	2,20
Encoderbl. 3	4,78	1,23	2,00	2,67
Encoderbl. 4	4,96	**1,46**	**2,38**	**2,93**
GRU	4,80	0,98	2,35	2,71
Multiplikation	2,70	–0,12	–0,12	0,82
Statisch (Enc. 4)	4,14	1,16	1,41	2,24

strumentensignalen. Darüber hinaus enthält die Tabelle 5.3 Ergebnisse für die Einspeisung von Konstanten als zusätzliche, statische Information über aktive Instrumente vor dem Encoderblock 4.

Insgesamt wird das durchschnittliche Separationsergebnis der Modelle mit drei Instrumenten durch die zusätzliche Informationseinspeisung um mindestens 1,44 dB gegenüber dem Fall ohne Zusatzeinspeisung gesteigert. Darüber hinaus ist das SI-SDR jedes der drei Instrumente für alle Einspeiseorte besser als das ohne Zusatzinformation. Folglich profitiert das Separationsmodell des Beispieltrios deutlich von der in dieser Arbeit vorgeschlagenen Integration der zeitabhängigen Instrumentenaktivität. Als beste Einspeiseposition wird auch hier die Einspeisung vor Encoderblock 4 bestätigt. Verglichen mit den Ergebnissen aus Tabelle 5.2 ist dabei vor allem das SI-SDR der Violine stark gestiegen. Ein Grund dafür ist die stärkere Fokussierung aller Merkmale auf die drei Instrumente des Beispieltrios, weshalb spezifische Eigenschaften der Violine im Training detaillierter gelernt werden. Wie oben für das größere Separationsmodell ausgeführt, verbessert die Multiplikation der zeitabhängigen Aktivität mit den geschätzten Instrumentensignalen auch das ursprüngliche Separationsergebnis des Modells mit drei Instrumenten nur marginal bzw. verschlechtert es sogar leicht im Falle der Violine. Des

Weiteren bestätigen sich die schlechteren Ergebnisse bei Einspeisung von statischer Zusatzinformation. Im exemplarisch dargestellten Fall vor Encoderblock 4 ist das SI-SDR im Durchschnitt um 0,69 dB verringert. Für andere Einspeisepositionen beträgt die Differenz zwischen zeitabhängiger und statischer Zusatzinformation ähnliche oder noch größere Werte, was den im Anhang zusammengestellten Tabellen B.4, B.5 und B.6 für SI-SDR, SI-SIR und SI-SAR aller gemeinsamen Separationsmodelle des definierten Beispieltrios entnommen werden kann.

In den Tabellen des Anhangs B ist in der letzten Spalte jeweils die Verbesserung der durchschnittlichen Metrik gegenüber einem Separationsmodell mit identischer Architektur aufgeführt, das nur mit dem nicht augmentierten URMP-Datensatz trainiert wurde. Diese Verbesserungen sind für alle betrachteten Modellarchitekturen positiv, wodurch der erfolgreiche Einsatz des in Abschnitt 5.3 beschriebenen, augmentierten Datensatzes untermauert wird. Die Steigerung des SI-SDR durch augmentierte Daten fällt für das Modell des definierten Beispieltrios mit über 4 dB im Allgemeinen höher aus als bei 13 Instrumenten. Das ist logisch, da die augmentierten Musikstücke nur aus Signalen dieser drei Instrumente zusammengesetzt sind.

Allen bisher diskutierten Separationsergebnissen mit Einspeisung von Zusatzinformation liegen fehlerfreie zeitabhängige Instrumentenaktivitäten zugrunde. Ist diese Information nicht vorab bekannt, muss sie geschätzt werden. Da die Schätzungen in der Regel Fehler enthalten, wird die Robustheit des vorgestellten Separationsansatzes gegenüber fehlerhaften Instrumentenaktivitäten anhand von simulativ erzeugten Fehlern untersucht. Dazu wird ein vorgegebener Prozentsatz der zusätzlich eingespeisten, binären Aktivitätswerte jedes Instruments invertiert, wobei die Auswahl der invertierten Werte zufällig erfolgt. Diese Robustheitsanalyse wird ausschließlich im Test der Separationsmodelle durchgeführt, weshalb die bereits mithilfe des fehlerfreien Datensatzes trainierten KNN zum Einsatz kommen und keine Trainingsdaten verfälscht werden müssen.

Abbildung 5.5 visualisiert den Zusammenhang zwischen durchschnittlichem SI-SDR des URMP-Testdatensatzes und dem Prozentsatz der zufällig invertierten Instrumentenaktivitäten für die Gesamtmodelle mit 13 Instrumenten und verschiedenen Einspeiseorten. Daraus geht hervor,

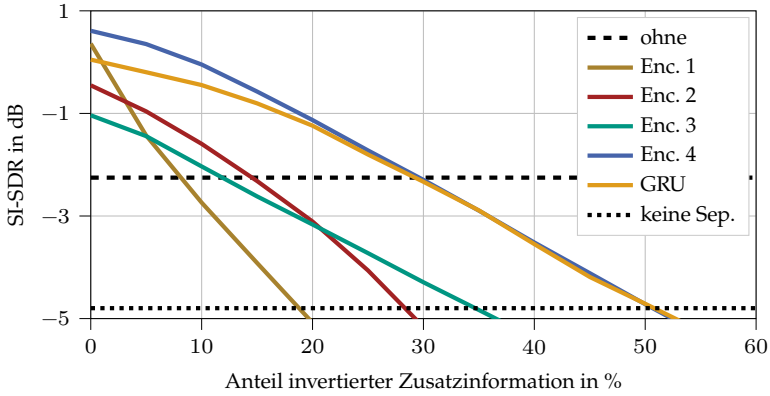


Abbildung 5.5 Durchschnittliche Separationsergebnisse des URMP-Testdatensatzes für gemeinsame Modelle mit 13 Instrumenten und unterschiedlichen Einspeisepositionen in Abhängigkeit von der zufälligen Invertierung eines Prozentsatzes der Zusatzinformation.

dass die beiden tiefsten Einspeisepositionen nahe den GRU-Modulen am robustesten sind, da sie bis zu einer Fehlerrate von ungefähr 30 % bessere Separationsergebnisse erzielen als ohne Zusatzinformation. Das verstärkt die bereits oben gezogene Schlussfolgerung, dass sich die Kombination von zeitabhängigen Instrumentenaktivitäten und abstrakten Merkmalen, die große Zeitspannen umfassen, am positivsten auf die Separation auswirkt. Die Einspeisung vor den ersten Encoderblöcken ist dagegen weniger robust, was in Abbildung 5.5 am deutlichsten anhand der Kurve für Encoderblock 1 zu sehen ist. Obwohl das SI-SDR bei perfekter Zusatzeinspeisung am zweitbesten ist, fällt es mit zunehmender Fehlerrate steil ab. Ein Grund dafür ist die starke Gewichtung der zusätzlichen Information mit 13 Kanälen am Modelleingang, welche das eindimensionale zu trennende Musiksignal in den Hintergrund rückt.

Die hohe Robustheit der Einspeiseposition vor Encoderblock 4 wird auch für den auf das Beispieltrio angepassten Fall bestätigt. In Abbildung 5.6 sind die zugehörigen Verlaufskurven des durchschnittlichen SI-SDR aller Einspeiseorte für die Separationsmodelle mit drei Instrumenten dargestellt. Dabei überzeugen vor allem die Separationsmodelle mit Zusatzeinspeisung vor Encoderblock 3 und 4 durch eine große Robustheit gegenüber den simulierten Fehlern. Im Vergleich zur Separation

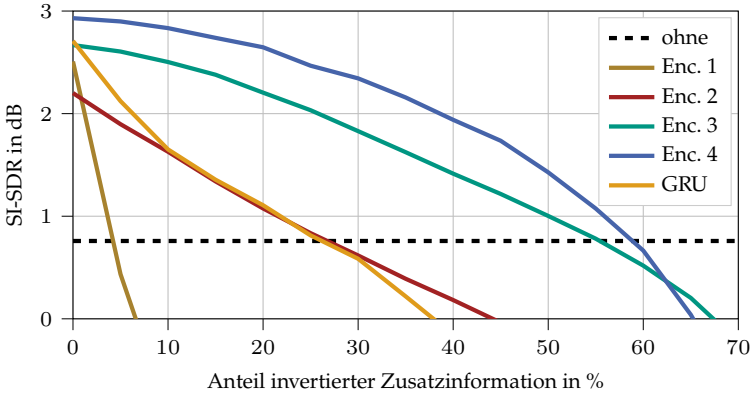


Abbildung 5.6 Durchschnittliche Separationsergebnisse der Instrumente des Beispieltrios für gemeinsame Modelle dieses Trios und unterschiedliche Einspeisepositionen in Abhängigkeit von der zufälligen Invertierung eines Prozentsatzes der Zusatzinformation.

ohne Zusatzinformation kann bei Einspeisung vor Encoderblock 4 sogar bis zu einem Fehleranteil von knapp 60 % eine Verbesserung erzielt werden. Ebenfalls bestätigt wird die eindeutig schlechteste Robustheit für die Einspeisung vor Encoderblock 1. Dagegen ist die Einspeiseposition vor der GRU für das Modell des Beispieltrios deutlich weniger robust als bei der Separation von 13 Instrumenten. Ein Grund ist der fehlende Encoderblock nach Einspeisung der Zusatzinformation, der die zeitabhängige Instrumentenaktivität mit den extrahierten Merkmalen verknüpft. Diese Verknüpfung ist bei wenigen Instrumenten relevanter, da die Beziehungen zwischen ihnen wichtiger sind.

Neben der Robustheitsanalyse des durchschnittlichen SI-SDR ermöglicht eine Untersuchung der instrumentenspezifischen Robustheit die Bewertung der Separation pro Instrument bei einer fehlerhaften Zusatzeinspeisung. Dazu ist der Verlauf der SI-SDR-Differenz zwischen den Ergebnissen des jeweiligen Trio-Instruments mit und ohne Zusatzinformation in Abbildung 5.7 visualisiert. Aus Gründen der Übersichtlichkeit wird ausschließlich der Einspeiseort vor Encoderblock 4 betrachtet, der die besten Ergebnisse erzielt. In Abbildung 5.7 werden sowohl das Separationsmodell mit 13 Instrumenten als auch das Modell des Beispieltrios analysiert. Die größte und robusteste Steigerung der Separationsergeb-

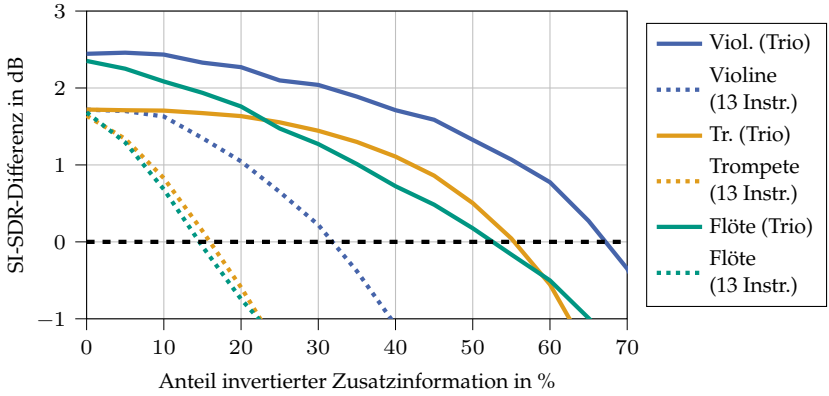


Abbildung 5.7 Instrumentenspezifische Verbesserung der Separationsergebnisse von Gesamtmodellen mit insgesamt 13 bzw. 3 Instrumenten durch Einspeisung der Instrumentenaktivität vor Encoderblock 4 im Vergleich zur Separation ohne Zusatzeinspeisung.

nisse wird für die Violine erzielt. Für die Separation aller 13 Instrumente fallen die Ergebnisse der Flöte und der Trompete deutlich schneller ab, weshalb ihre Robustheit geringer ist. Im Falle des auf das Beispieltrio spezialisierten Modells sind die Ergebnisse dagegen nur leicht schlechter als die der Violine. Sie alle übertreffen das SI-SDR der Separation ohne Zusatzinformation bis mindestens zu einer Fehlerrate der zusätzlichen Instrumentenaktivität von über 50 %, was die Robustheit der Einspeisung vor Encoderblock 4 unterstreicht. Insgesamt steigen sowohl die instrumentenspezifischen Ergebnisse als auch die Robustheit, wenn das gemeinsame Separationsmodell auf eine definierte Besetzung wie das Beispieltrio aus Violine, Trompete und Flöte trainiert wird.

5.5.2 Unabhängige Separationsmodelle

Als Alternative zu den im vorangegangenen Abschnitt untersuchten, gemeinsamen Gesamtmodellen mit I_{Instr} Ausgangssignalen werden im Folgenden die schon in Abschnitt 5.2 beschriebenen, unabhängigen Separationsmodelle analysiert. Sie extrahieren aus der eingespeisten Ensemble-Aufnahme jeweils nur das Quellensignal eines definierten Instruments und behandeln die übrigen Instrumentenanteile implizit als Störgeräu-

sche. Darüber hinaus wird nur die zeitabhängige Instrumentenaktivität des entsprechenden Instruments eingespeist, wenn die Separation mit Zusatzinformation ausgeführt ist. Durch die Unabhängigkeit von anderen Signalquellen können beliebig viele Separationsmodelle verschiedener Instrumente flexibel hinzugefügt, ersetzt oder weggelassen werden. Dies kann auch noch zu einem späteren Zeitpunkt erfolgen, sodass die Separation je nach zu trennender Besetzung adaptiert werden kann. Des Weiteren beeinflussen sich die Ergebnisse der einzelnen Modelle nicht.

Die unabhängigen KNN der I_{Instr} zu trennenden Instrumente, die jeweils kleiner als ein gemeinsames Separationsmodell sind, werden zur Separation von Ensemble-Aufnahmen als ein Separationssystem aufgefasst. Jedes Modell des Systems besitzt dabei eine identische Architektur. Nur das zu extrahierende Instrument ist verschieden. Die Fokussierung auf dieses Instrument wird während des Trainings durch die Anpassung des gewünschten Instrumentensignals am Ausgang und die Anpassung der Zusatzinformation realisiert. Ansonsten erfolgt das Training jedes Modells gleich und auf Basis des in Abschnitt 5.3 definierten, augmentierten URMP-Datensatzes. Der Einfluss des Einspeiseortes wird mithilfe von Separationsmodellen aller möglichen Einspeisepunkte im Encoder untersucht. Zur Bewertung der Separationsqualität werden, analog zu den Betrachtungen in Abschnitt 5.5.1, nur die sechs im URMP-Testdatensatz vorhandenen Musikinstrumente analysiert, da die geschätzten Signale der übrigen Instrumente fast ausschließlich Stille enthalten.

Für die Separationssysteme mit Zusatzinformation an unterschiedlichen Einspeiseorten ergeben sich damit die in Tabelle 5.4 aufgelisteten SI-SDR-Werte. Die Ergebnisse des Modells ohne Zusatzeinspeisung (-), der direkten Übertragung des Eingangssignals an alle Ausgänge (‚Mix‘) und der Multiplikation von den ohne Zusatzinformation geschätzten Quellensignalen mit der zugehörigen Instrumentenaktivität (‚Mult.‘) dienen dabei zur Einordnung der Separationsqualität. Als weitere Referenz sind die Ergebnisse des Literaturansatzes Open-Unmix [132] angegeben, welcher auf Zeit-Frequenz-Darstellungen aller Musiksignale basiert und ebenfalls je ein KNN pro zu trennendem Instrument verwendet. Um einen fairen Vergleich zu ermöglichen und die Separation der identischen Musikinstrumente durchzuführen, wurde das Open-Unmix-System mit dem augmentierten URMP-Datensatz neu trainiert.

Tabelle 5.4 SI-SDR (in dB) aller Instrumente des URMP-Testdatensatzes für unabhängige Separationsmodelle ohne (-) und mit Zusatzeinspeisung an unterschiedlichen Einspeiseorten. Als Referenzen dienen das nicht separierte Eingangssignal (‚Mix'), die Multiplikation der zeitabhängigen Instrumentenaktivität mit der Schätzung ohne sie (‚Mult.') und der Literaturansatz Open-Unmix [132].

	Fagott	Flöte	Oboe	Sax.	Tromp.	Violine	∅
-	–44,81	1,99	–2,28	–14,81	3,38	–1,00	–9,59
Enc. 1	–5,47	0,46	–0,98	–6,21	1,58	–2,36	–2,16
Enc. 2	–6,69	1,63	–0,72	–3,98	1,22	0,10	–1,41
Enc. 3	–7,05	2,42	**–0,35**	**–3,39**	4,03	0,42	–0,65
Enc. 4	–6,63	1,90	–0,98	–4,76	4,21	0,69	–0,93
GRU	–4,00	**3,86**	–1,32	–4,78	**4,38**	0,68	**–0,20**
Mix	**–3,51**	1,29	–11,29	–6,17	–4,12	–4,98	–4,80
Mult.	–44,54	2,12	–2,28	–14,77	3,46	–1,05	–9,51
[132]	–7,52	1,92	–0,45	–30,92	0,08	**2,53**	–5,73

Anhand der durchschnittlichen Separationsergebnisse in Tabelle 5.4 wird deutlich, dass die zusätzliche Einspeisung der zeitabhängigen Instrumentenaktivität zu einer verbesserten Separation von Ensemble-Aufnahmen durch Einzelmodelle führt. Für jede Einspeiseposition übertrifft das durchschnittliche SI-SDR die Werte aller angegebenen Referenzen. Die besten Separationsergebnisse werden bei Einspeisung direkt vor dem GRU-Modul erzielt, gefolgt von den Einspeisungen vor Encoderblock 3 und 4. Das bestätigt die Ergebnisse mit einem gemeinsamen Modell aus Abschnitt 5.5.1, wonach der Mehrwert der Zusatzeinspeisung bei Kopplung mit abstrakten Merkmalen tief in der UNet-Architektur am größten ist. Grundsätzlich gilt diese Tendenz auch für die Einzelergebnisse der Instrumente. Es ist allerdings kein allgemeingültiger Zusammenhang erkennbar. So bleiben z. B. alle SI-SDR-Werte für das Fagott unter dem Resultat des nicht separierten Eingangssignals zurück. Dieser Spezialfall ist auf die geringe Menge an Trainingsdaten für das Fagott zurückzuführen, wegen der keine ausreichenden, charakteristischen Merkmale des Fagotts gelernt werden können. Folglich ist eine Extraktion der Fagottaufnahme nicht zufriedenstellend, was vor allem an dem extremen SI-SDR für die

Separation ohne Zusatzinformation deutlich wird. Ähnliches gilt auch für das Saxofon, das zudem aufgrund seiner großen Klangvariabilität schwieriger zu extrahieren ist.

Eine Multiplikation der zeitabhängigen Instrumentenaktivität am Ausgang verbessert die Ergebnisse ohne Zusatzeinspeisung wie im Falle der gemeinsamen Separationsmodelle nur marginal bzw. verschlechtert sie sogar leicht im Falle der Violine. Daraus folgt, dass auch die unabhängigen Einzelmodelle kaum Signalanteile in den nicht aktiven Passagen des Instruments schätzen, was sehr positiv ist. Der Literaturansatz Open-Unmix, der ebenfalls keine Zusatzinformation nutzt, erzielt bessere Separationsergebnisse als das modifizierte Demucs-Modell dieser Arbeit ohne Zusatzinformation. Ausnahmen davon sind das Saxofon und die Trompete. Mit Zusatzeinspeisung können die Ergebnisse von Open-Unmix dagegen für die meisten Instrumente übertroffen werden. Die Ausnahme stellt hier die Violine dar, welche durch Open-Unmix wesentlich besser extrahiert wird. Sie spielt in den Testdaten meist die Melodie, sodass sie mehr im Vordergrund zu hören ist, was einer auf Zeit-Frequenz-Darstellungen basierenden Separation offensichtlich zugute kommt. Darüber hinaus besitzt auch die Violine eine breite Klangvariabilität, die von einer Matrixmultiplikation mit der ursprünglichen Eingangsdarstellung umfänglicher extrahiert werden kann. Um das Separationsergebnis zu optimieren, könnte theoretisch eine beliebige Kombination aus unabhängigen Einzelmodellen mit unterschiedlichen Separationsmethoden und Zusatzeinspeisungen gewählt werden. Im Folgenden wird aber immer eine Methode für alle Instrumente gewählt, um einen eindeutigen Vergleich vornehmen zu können.

Die Auswirkung der Zeitabhängigkeit der Zusatzinformation wird exemplarisch am Separationssystem mit Einspeisung vor dem GRU-Modul analysiert, da es das im Durchschnitt beste Separationsergebnis liefert. Das System wird einmal wie im Training mit der zeitabhängigen Instrumentenaktivität gespeist und im anderen Fall mit einem statischen Vektor aus Einsen oder Nullen versorgt, um einen einzelnen Wert zur Präsenz des Instruments im Musikstück zu imitieren, wie er z. B. im Literaturansatz von Slizovskaia et al. [126] verwendet wird. In Abbildung 5.8 sind die resultierenden Separationsergebnisse grafisch dargestellt. Daraus wird ersichtlich, dass die Flöte, das Saxofon und die Trompete von

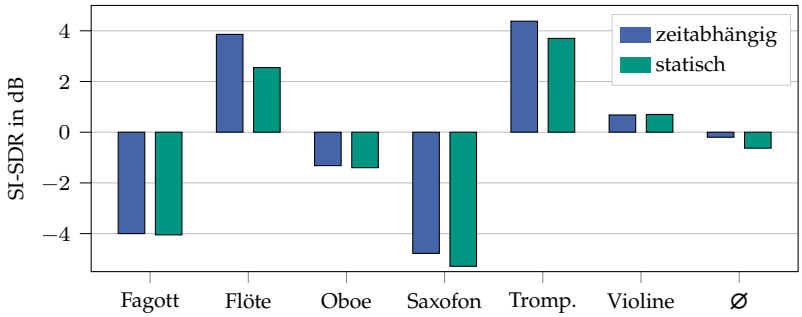


Abbildung 5.8 Vergleich der Separationsergebnisse des URMP-Testdatensatz bei zeitabhängiger und statischer Information über aktive Instrumente, die in den unabhängigen Einzelmodellen vor dem ersten GRU-Modul eingespeist wird.

der Zeitabhängigkeit der Zusatzinformation merklich profitieren. Die Separation der anderen drei Instrumente verbessert sich dagegen nur marginal oder bleibt annähernd gleich, sodass die Zeitabhängigkeit der Instrumentenaktivität für sie keinen großen Einfluss hat. Im Durchschnitt steigert sich das SI-SDR um 0,43 dB, wodurch die bessere und gezieltere Separation durch eine zeitabhängige Zusatzinformation unterstrichen wird. Für die anderen Einspeiseorte ergeben sich vergleichbare oder höhere Verbesserungen durch die Zeitabhängigkeit.

Alle Zahlenwerte für jede Einspeiseposition der zeitabhängigen und der statischen Zusatzinformation sind nochmal ausführlich in den Tabellen B.7, B.8 und B.9 aufgelistet. Sie enthalten alle SI-SDR-, SI-SIR- und SI-SAR-Werte der bisher diskutierten, unabhängigen Separationsmodelle sowie ihre jeweilige Verbesserung durch die Augmentierung des URMP-Datensatzes. Diese Verbesserungen durch die Datenaugmentierung fallen für die Systeme mit Einzelmodellen geringer aus als für die gemeinsamen Separationsmodelle. Das liegt an den Aktivitätsinformationen aller betrachteten Instrumente, welche in den Gesamtmodellen während jeder Separation zur Verfügung stehen. Dadurch sind alle Musikstücke des Datensatzes relevante Trainingsdaten für die Gesamtmodelle, wohingegen die Separation durch Einzelmodelle im Training hauptsächlich durch Musikstücke verbessert werden kann, in denen das relevante Musikinstrument spielt. Folglich ist die Augmentierung fast ausschließlich

Tabelle 5.5 SI-SDR (in dB) der Instrumente des Beispieltrios für unabhängige Separationsmodelle ohne (-) und mit Zusatzeinspeisung an unterschiedlichen Orten. Als Referenzen dienen die Multiplikation der Zusatzinformation am Ausgang, die Einspeisung statischer Instrumentenaktivität vor dem GRU-Modul und der Literaturansatz Open-Unmix [132].

	Flöte	Tromp.	Violine	∅
-	1,99	3,38	−1,00	1,45
Encoderbl. 1	0,46	1,58	−2,36	−0,11
Encoderbl. 2	1,63	1,22	0,10	0,98
Encoderbl. 3	2,42	4,03	0,42	2,29
Encoderbl. 4	1,90	4,21	0,69	2,26
GRU	**3,86**	**4,38**	0,68	**2,97**
Multiplikation	2,12	3,46	−1,05	1,51
Statisch (GRU)	2,55	3,70	0,70	2,31
Op.-Unm. [132]	1,92	0,08	**2,53**	1,51

für die unabhängigen Einzelmodelle der Violine, der Trompete und der Flöte von Vorteil, weshalb dieses Beispieltrio im Folgenden nochmal gesondert betrachtet wird.

Die Ergebnisse des Separationssystems aus den drei Einzelmodellen des Beipieltrios sind in Tabelle 5.5 für alle Einspeiseorte zusammengestellt. Zusätzlich sind die Separationsergebnisse der Multiplikation von den ohne Zusatzinformation geschätzten Signalen mit der zeitabhängigen Instrumentenaktivität, die statische Zusatzeinspeisung vor dem GRU-Modul und die Ergebnisse des mit dem augmentierten URMP-Datensatz trainierten Open-Unmix-Modells angegeben. Ohne die sehr schlechten Ergebnisse für Fagott und Saxofon (s. Tabelle 5.4) ist die durchschnittliche Separation der Einzelmodelle ohne Zusatzinformation vergleichbar zu dem Ergebnis des Literaturansatzes Open-Unmix, welcher ebenfalls keine Zusatzinformation verwendet. Sie unterscheiden sich aber sehr deutlich im SI-SDR von Trompete und Violine, die jeweils über 3 dB Differenz aufweisen. Die in dieser Arbeit vorgeschlagene Modellarchitektur extrahiert dabei die Trompete wesentlich besser und Open-Unmix die Violine. Dieser Unterschied lässt sich durch die beiden Konzepte der Separation erklären, da Zeit-Frequenz-Darstellungen die

Trennung des Melodieinstruments im Vordergrund oft erleichtern, das in diesem Fall die Violine darstellt.

Durch Integration der zeitabhängigen Instrumentenaktivität lassen sich die Separationsergebnisse nur verbessern, wenn die zusätzliche Information vor Encoderblock 3, Encoderblock 4 oder dem GRU-Modul eingespeist wird. Dort steigert die Zusatzeinspeisung das durchschnittliche SI-SDR um mindestens 0,81 dB. Damit wird die oben ausgeführte Schlussfolgerung bestätigt, dass die Separation von der Kombination der Zusatzinformation mit abstrakten Merkmalen, die einen größeren Zeitbereich umfassen, profitiert. Die beste Separation wird bei Einspeisung vor dem GRU-Modul erzielt. Frühe Einspeiseorte wie vor Encoderblock 1 oder 2 sind dagegen nicht geeignet, da sie insgesamt und vor allem für die Trompete schlechtere Separationsergebnisse liefern als ohne Zusatzinformation. Dies kann darauf hindeuten, dass die zugehörigen KNN die charakteristischen Merkmale der Instrumente im Training nicht gut gelernt haben und sie sich in der Signalschätzung zu stark auf die Zusatzinformation verlassen. Für die Multiplikation der Instrumentenaktivität und die statische Zusatzeinspeisung vor dem GRU-Modul werden die bereits diskutierten Zusammenhänge bestätigt.

Im Vergleich zu den besten gemeinsamen Separationsmodellen des Beispieltrios in Tabelle 5.3 ist das SI-SDR der besten Einzelmodelle der Trompete in Tabelle 5.5 stark erhöht. Daran wird deutlich, dass die gemeinsamen Modelle sich weniger auf die Separation der Trompete und ihre charakteristischen Eigenschaften fokussieren, sodass die unabhängige Extraktion des Trompetensignals mithilfe eines Einzelmodells Vorteile bringt. Der Fokus der gemeinsamen Modelle liegt dagegen mehr auf der Flöte und der Violine, weshalb deren Ergebnisse deutlich über denen der Einzelmodelle liegen. Gerade die Separation der Violine profitiert stark von den Aktivitätsverläufen der beiden anderen Instrumente. Insgesamt verzeichnen die besten Varianten beider Modellarchitekturen mit 2,93 dB bzw. 2,97 dB aber fast identische Durchschnittswerte des SI-SDR, sodass keine Architektur überlegen ist.

Auch für die Separationssysteme aus unabhängigen Einzelmodellen wird eine Robustheitsanalyse durchgeführt, um die Separationsqualität bei fehlerhaften Zusatzeinspeisungen abschätzen zu können. Dafür wird ein vorgegebener Prozentsatz der bekannten Instrumentenaktivitäten

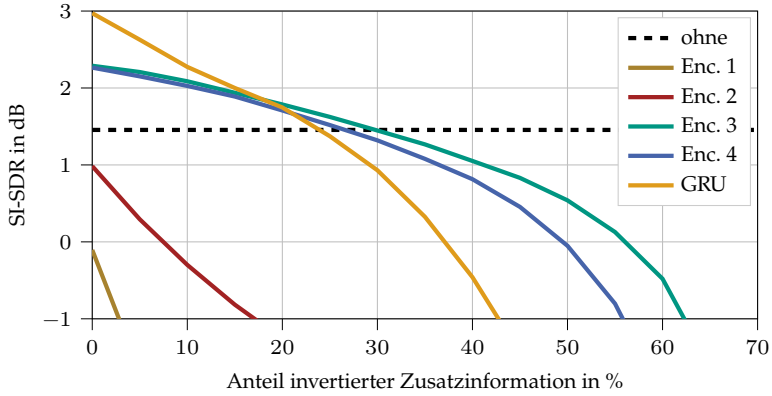


Abbildung 5.9 Durchschnittliche Separationsergebnisse der Instrumente des Beispieltrios für unabhängige Einzelmodelle dieses Trios und unterschiedliche Einspeisepositionen in Abhängigkeit von der zufälligen Invertierung eines Prozentsatzes der Zusatzinformation.

des URMP-Testdatensatzes zufällig invertiert. Die daraus resultierenden Kennlinien der durchschnittlichen SI-SDR-Werte sind für die Separationsmodelle des Beispieltrios sowie alle Einspeiseorte in Abbildung 5.9 dargestellt. Es wird nur der Fall des Beispieltrios betrachtet, da seine durchschnittlichen SI-SDR-Werte prinzipiell ähnliche Verläufe wie für den Fall mit allen sechs Instrumenten des URMP-Testdatensatzes zeigen, aber nicht von einzelnen sehr schlechten Ergebnissen für selten im Trainingsdatensatz vorkommende Instrumente (z. B. Fagott oder Saxofon) überlagert werden.

Die Separationsmodelle mit Zusatzeinspeisung vor Encoderblock 1 und 2 erzielen die schlechtesten Ergebnisse und sind auch aufgrund ihrer fehlenden Robustheit ungeeignet. Am robustesten sind dagegen die Einspeisepositionen vor Encoderblock 3 und 4, da die jeweiligen Kennlinien bei größer werdender Fehlerrate am flachsten abfallen. Das stimmt mit der hohen Robustheit der entsprechenden gemeinsamen Separationsmodelle überein. Im Falle der hier analysierten, unabhängigen Einzelmodelle sind die extrahierten Instrumentensignale mit Zusatzeinspeisung vor dem GRU-Modul am besten, weil die durchschnittliche Separationsqualität dieser Modelle bis zu etwa 20 % zufälliger Fehler am höchsten ist. Den Fall ohne Zusatzeinspeisung übertrifft das Separations-

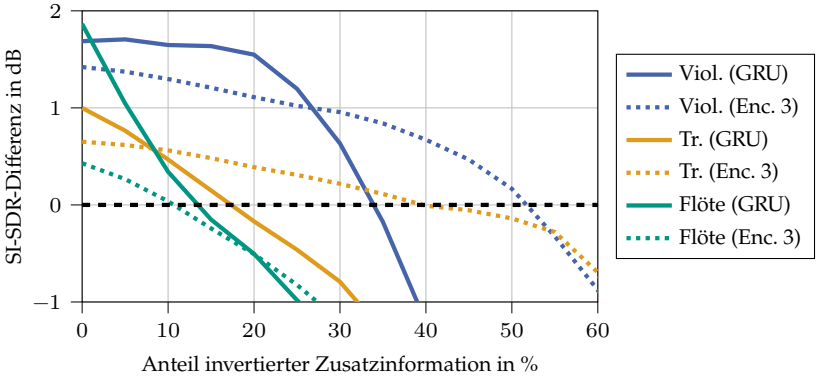


Abbildung 5.10 Instrumentenspezifische Verbesserung der Separationsergebnisse von Einzelmodellen durch die Einspeisung der zeitabhängigen Instrumentenaktivität an den beiden besten Einspeisepositionen im Vergleich zur Separation ohne Zusatzeinspeisung.

system mit Einspeisung vor dem GRU-Modul bis zu einer Fehlerrate von ungefähr 23 %. Verglichen mit den gemeinsamen Separationsmodellen in Abbildung 5.6 sind die Systeme aus unabhängigen Modellen weniger robust, da sie über keine zusätzlichen Informationen der anderen Instrumente verfügen, mithilfe derer fehlerhafte Instrumentenaktivitäten zu einem gewissen Teil aufgefangen werden können.

Für die beiden besten Einspeisepositionen der Einzelmodelle, Encoderblock 3 und GRU, werden auch die Verläufe der instrumentenspezifischen Robustheit von Violine, Trompete und Flöte untersucht. Sie sind in Abbildung 5.10 anhand der SI-SDR-Differenz zum Fall ohne Zusatzeinspeisung und in Abhängigkeit der simulierten Fehlerrate dargestellt. Die Separationsergebnisse der Flöte nehmen in beiden Fällen schon bei wenigen zufälligen Fehlern stark ab, sodass die Ergebnisse bereits bei einer Fehlerrate von etwas mehr als 10 % schlechter sind als im Fall ohne Zusatzinformation. Bei Einspeisung vor dem GRU-Modul ist die Robustheit der Trompete nur leicht besser als die der Flöte, wohingegen die Extraktion der Violine bis ungefähr 20 % zufällig invertierter Zusatzinformation annähernd gleich gut ist und bei mehr Fehlern dann drastisch abfällt. Eine Einspeisung vor Encoderblock 3 führt dagegen zu einer robusteren Separation von Trompete und Violine, da die entsprechenden

Kurven in Abbildung 5.10 mit zunehmendem Fehler nur langsam sinken. Dies unterstreicht die robustere Integration der Zusatzinformation vor einem tiefen Encoderblock. Trotzdem sind die absoluten SI-SDR-Werte bei Einspeisung vor dem GRU-Modul besser, falls die Fehlerrate der Instrumentenaktivität niedrig ist.

5.6 Auswirkungen realer Instrumentendetektionsergebnisse als Zusatzinformation

Die Untersuchungen in Abschnitt 5.5 zeigen, dass die in dieser Arbeit vorgeschlagene Zusatzeinspeisung der zeitabhängigen Instrumentenaktivität die Separation von Ensemble-Aufnahmen unterschiedlicher Instrumententypen verbessert. In den bisherigen Experimenten kommen dabei stets die wahren, vorab bekannten Instrumentenaktivitäten zum Einsatz, die entweder direkt oder durch zufällige, simulierte Fehler manipuliert in das Separationsmodell eingespeist werden. Sofern diese Aktivitäten vorab bekannt sind, können sie auch in der realen Anwendung als Zusatzinformation genutzt werden. Das ist in den meisten Fällen aber nicht der Fall, sodass eine vorgeschaltete Schätzung der aktiven Instrumente notwendig ist. Ein geeigneter Ansatz dazu ist die in Kapitel 4 vorgestellte zeitabhängige Instrumentendetektion, deren nicht perfekte Schätzung in diesem Abschnitt zur Verifizierung der Separationsergebnisse mit simulierter Zusatzinformation dient.

Um die Zusatzinformation für möglichst viele Instrumente zu schätzen und damit einen allgemeineren Anwendungsfall zu analysieren, wird eine Variante des Detektionsmodells ID_ZF_Res mit allen 11 Instrumenten des MusicNet-Datensatzes [137] trainiert. Aufgrund der besten Ergebnisse in Abschnitt 4.4 wird das KNN mit einer CQT des zu analysierenden Musikabschnitts gespeist. Die Zeitdimension der CQT ist mit 128 Zeitbins kleiner als in den Experimenten, weil sie an die Eingangsgröße des Separationsmodells von 65 536 Zeitwerten angepasst ist. Dadurch kann die Instrumentendetektion und die anschließende Separation mit den gleichen Musiksequenzen der Länge 65 536 erfolgen und theoretisch auch schon während der Aufnahme durchgeführt werden, sobald eine

Sequenz vollständig aufgenommen ist. Auf Basis jeder CQT der Dimension $(400, 128)$ schätzt das modifizierte Detektionsmodell am Ausgang 32 Zeitwerte für alle Instrumente. Diese Zahl resultiert aus den in Tabelle 4.1 aufgeführten Parametern für die Dimension $(400, 240)$, wobei die Größe d_{MP4} des letzten *Max-Poolings* auf (4×1) reduziert ist. Folglich besitzt die geschätzte Instrumentendetektion eine höhere Zeitauflösung von ca. 46,44 ms als die in Abschnitt 4.4 untersuchten Modellvarianten, was der hohen zeitlichen Auflösung des auf Zeitsignalen basierenden Separationsmodells entgegenkommt. Der Nachteil dieser höheren Zeitauflösung ist die damit verbundene, reduzierte Detektionsgüte von im Durchschnitt 87,30 % für das F-Maß des in den Experimenten zur zeitabhängigen Instrumentendetektion verwendeten MusicNet-Testdatensatzes. Das entspricht einer Reduktion von 1,52 % gegenüber dem besten Modell aus Kapitel 4, was angesichts der größeren Zahl an geschätzten Instrumenten vertretbar ist.

Im Training des modifizierten Modells zur zeitabhängigen Instrumentendetektion wird der MusicNet-Datensatz genutzt, da er größer als der URMP-Datensatz ist. Die Abtastfrequenz der MusicNet-Aufnahmen ist 44,1 kHz, sodass die geschätzten Instrumentenaktivitäten zur Einspeisung in das Separationsmodell auf die Abtastfrequenz des URMP-Datensatzes von 48 kHz gebracht werden müssen. Auch eine Schätzung auf Basis der CQT-Darstellung mit 48 kHz Abtastfrequenz ist möglich, da die F-Maße der daraus resultierenden Aktivitäten weniger als 0,5 % von den Ergebnissen der Neuabtastung der mit 44,1 kHz geschätzten Aktivitäten abweichen. Trotzdem wird die Instrumentendetektion in den folgenden Experimenten bei der im Training verwendeten Abtastfrequenz von 44,1 kHz vorgenommen. Danach werden die Aktivitäten binarisiert, mithilfe eines Kaiser-Fensters neu abgetastet und anschließend nochmal auf die Werte 0 oder 1 quantisiert. Zur Binarisierung der Instrumentenaktivität werden verschiedene instrumentenspezifische Schwellenwerte verwendet, um unterschiedliche Detektionsgüten abzubilden. Dies ermöglicht die Analyse der Beziehung zwischen der Güte der Zusatzinformation und den resultierenden Separationsergebnissen.

Die binären, neu abgetasteten Aktivitätssignale jedes zu separierenden Instruments werden in die besten Separationsmodelle aus Abschnitt 5.5 eingespeist, um deren Robustheit hinsichtlich echter Schätzungen zu

analysieren. Leider stimmen die Instrumente des MusicNet- und des URMP-Datensatzes nur teilweise überein, sodass die Aktivität mancher Instrumente nicht mit dem verwendeten Detektionsansatz geschätzt werden kann. Deshalb werden für diese Instrumente die zwei Extremfälle der Einspeisung von perfekter, vorab bekannter Zusatzinformation und der Einspeisung des Nullvektors als schlechteste Aktivitätsschätzung eines aktiven Instruments untersucht.

Exemplarisch werden die Separationsergebnisse des zuvor definierten Beispieltrios aus Violine, Trompete und Flöte analysiert, die bei Einspeisung der echten Schätzungen für die zeitabhängige Instrumentenaktivität auftreten. Für die Violine und die Flöte können Schätzungen unterschiedlicher Güte generiert werden, die Trompete ist dagegen nicht im MusicNet-Datensatz enthalten und kann deshalb nicht geschätzt werden. Die Ergebnisse mit real geschätzten Zusatzinformationen sind für das gemeinsame Separationsmodell des Trios mit Einspeisung vor Encoderblock 4 in Abbildung 5.11 durch Punkte und Kreuze illustriert. Zur Einordnung dieser Ergebnisse sind darüber hinaus die Verläufe von Violine und Flöte aus Abbildung 5.7 dargestellt, welche den Effekt von simulierten, zufälligen Fehlern in der zusätzlichen Information veranschaulichen. Wichtig zum Verständnis der Ergebnisse ist dabei, dass das eingesetzte Separationsmodell das gleiche wie in Abschnitt 5.5.1 ist und mit der wahren Zusatzinformation des URMP-Datensatzes trainiert wurde. Nur im Test werden echte Schätzungen eingespeist. Somit können im Separationsmodell sowohl vorab bekannte als auch von einem beliebigen Ansatz geschätzte Instrumentenaktivitäten integriert werden.

Im Vergleich zu den simulierten Fehlern durch zufällige Invertierung führt die Einspeisung der geschätzten Instrumentenaktivitäten zu deutlich weniger robusten Separationsergebnissen, da alle einzelnen Punkte in Abbildung 5.11 unterhalb der Kennlinien liegen und mit zunehmender Fehlerrate weiter von ihnen entfernt sind. Das liegt an den in realen Schätzungen meist zusammenhängend auftretenden Fehlern, durch welche die Präsenz der Instrumente über einen längeren Zeitraum nicht korrekt geschätzt wird. Ursachen dafür sind z. B. nicht erkannte Noten bzw. Passagen eines Instruments, untypische bzw. nicht in den Trainingsdaten erlernte Instrumentenklänge, eine zeitliche Verschiebung der Detektion (Jitter) oder ungeeignete Schwellenwerte zur Binarisierung. Durch

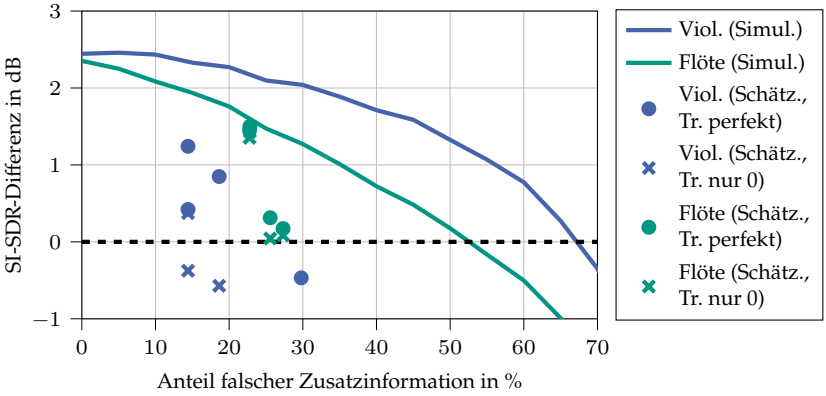


Abbildung 5.11 Verbesserung der Separationsergebnisse des Trio-Gesamtmodells durch Einspeisung von geschätzter und simulierter Instrumentenaktivität vor Encoderblock 4 im Vergleich zur Separation ohne Zusatzeinspeisung. Parallel zu den Schätzungen der zeitabhängigen Instrumentendetektion von Violine und Flöte wird für die Trompete entweder die perfekte Aktivität oder nur der Nullvektor eingespeist.

die zusammenhängenden Fehler ist das Separationsmodell nicht in der Lage, einzelne fehlerhafte Werte auf Basis von Nachbarinformationen zu vernachlässigen. Folglich sind die getrennten Instrumentensignale schlechter als im Falle von zufälligen Invertierungen.

Dennoch kann die Separation ohne Zusatzinformation bei einer ausreichend guten Instrumentendetektion durch geeignete Schwellenwerte übertroffen werden. Für die in Abbildung 5.11 dargestellten Ergebnisse des URMP-Testdatensatzes gilt das insbesondere für die Flöte, welche für zwei Konfigurationen der Detektion eine Fehlerrate von ungefähr 22,8 % aufweist und unabhängig von der eingespeisten Aktivität der Trompete eine Verbesserung von rund 1,4 dB erzielt. Im Falle der Violine ist die Separation auf Basis von echten Schätzungen der Instrumentenaktivität schwieriger, da ihr Separationsergebnis auch von der Zusatzeinspeisung für die Trompete abhängt. Diese Kopplung wird an der großen Distanz zwischen Punkten und Kreuzen der gleichen Fehlerrate ersichtlich. Das Kreuz für knapp 30 % liegt beispielsweise bei −3,28 dB und somit deutlich außerhalb des dargestellten Wertebereichs. Trotzdem erzielt die beste Konfiguration bei 14,42 % Fehlerrate eine Verbesserung von 0,37 dB

gegenüber dem Fall ohne Zusatzinformation, obwohl für die Trompetenaktivität der Nullvektor eingespeist wird. Bei einer realistischen Trompetendetektion zwischen perfekter Information und Nullvektor ist zu erwarten, dass ein höheres SI-SDR als mit Nullvektor erzielt wird. Damit ist die Verbesserung der Separation von Ensemble-Aufnahmen auch bei Einspeisung von zeitabhängigen Instrumentenaktivitäten bestätigt, welche durch einen unabhängigen Ansatz zur Instrumentendetektion geschätzt werden. Dabei ist eine hohe Detektionsgüte essenziell, die in diesem Fall durch geeignete Schwellenwerte realisiert wird.

Die Abhängigkeit der Separationsergebnisse eines Musikinstruments von der Zusatzinformation eines anderen Instruments ist im gemeinsamen Separationsmodell teilweise sehr stark ausgeprägt. Das birgt die Gefahr einer negativen Beeinflussung der Separation durch fehlerbehaftete Zusatzinformationen anderer Quellen. Darüber hinaus kann der in der Simulation ermittelte Zusammenhang zwischen fehlerhafter Instrumentenaktivität und den daraus resultierenden Separationsergebnissen nicht für jedes Instrument isoliert verifiziert werden. Mithilfe der alternativen Modellarchitektur mit unabhängigen Separationsmodellen jedes Instruments kann diese instrumentenspezifische Betrachtung durchgeführt werden, ohne dass Zusatzinformationen anderer Instrumente eine Rolle spielen. Dazu sind in Abbildung 5.12 exemplarisch die Separationsergebnisse der Einzelmodelle von Violine und Flöte mit Einspeisung von simulierter und geschätzter Instrumentenaktivität vor dem GRU-Modul dargestellt. Für die Instrumentendetektion werden sehr unterschiedliche Schwellenwerte untersucht, um einen großen Bereich der Fehleranteile abzudecken. Die kleinsten Fehlerraten von 13,74 % für die Violine und 22,33 % für die Flöte führen dabei zu SI-SDR-Werten von 0,06 dB bzw. 2,45 dB, die jeweils besser als die in Tabelle 5.5 angegebenen Ergebnisse ohne Zusatzeinspeisung sind. Folglich kann die Separation mit Einzelmodellen durch Einspeisung von ausreichend guten Aktivitätsschätzungen der zu trennenden Instrumente verbessert werden.

Im Gegensatz zu den Ergebnissen des gemeinsamen Separationsmodells in Abbildung 5.11 werden die simulativ erzeugten Kennlinien für Violine und Flöte durch die echten Schätzungen validiert, da die Punkte der realen Schätzungen in Abbildung 5.12 den groben Verlauf der Kennlinien nachbilden. Der Hauptgrund dafür ist die Fokussierung jedes

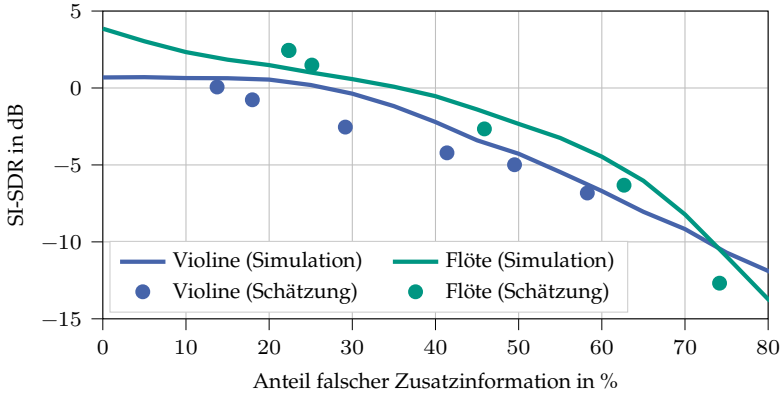


Abbildung 5.12 Separationsergebnisse der unabhängigen Einzelmodelle für Violine und Flöte bei Zusatzeinspeisung von geschätzter und simulierter Instrumentenaktivität vor dem GRU-Modul.

unabhängigen Einzelmodells auf ein Instrument, wodurch Beeinflussungen der Separation durch Zusatzinformationen anderer Instrumente ausgeschlossen werden. Zwischen Fehleranteilen der Zusatzinformation von 15 % und 50 % sind die SI-SDR-Werte der Violine etwas niedriger als bei zufälliger Invertierung, was auf die stärkere Auswirkung von zusammenhängenden Fehlern im Separationssignal zurückzuführen ist. Bis auf diese kleine Abweichung passen die Ergebnisse mit echten Schätzungen aber sehr gut zu den simulierten Kennlinien. Folglich ist es im Falle der unabhängigen Einzelmodelle möglich, den Zusammenhang zwischen Fehlerrate und zugehörigen SI-SDR-Werten in der Simulation anhand von zufälligen Invertierungen abzuschätzen.

6 Fazit

6.1 Zusammenfassung

In dieser Arbeit wurde die Analyse polyphoner Musikaufnahmen auf drei unterschiedlichen Themengebieten weiterentwickelt. Zunächst wurde eine objektive Bewertung der Aufnahmequalität mithilfe von drei Parametern vorgeschlagen, die von kleinen, performanten KNN ohne Kenntnis der Musikaufnahme geschätzt werden. Der erste Qualitätsparameter ist das SNR einer 30-sekündigen Aufnahme bei statischem Hintergrundrauschen, das für vier Rauscharten mit einer durchschnittlichen Abweichung von unter 1 dB sehr zuverlässig geschätzt wird. Kurze, über die Zeit veränderliche Störgeräusche werden mithilfe einer zeitabhängigen SNR-Schätzung über Abschnitte von 2 s erfasst. Dieser zweite Qualitätsparameter wird ebenfalls zuverlässig mit einer durchschnittlichen Abweichung von rund 2,3 dB geschätzt. Als dritte Parameterklasse wurden die Nachhall- und die Anfangsnachhallzeit der 30-sekündigen Aufnahme betrachtet. Sie ermöglichen eine einfache Beschreibung der Raumakustik und werden im verwendeten Datensatz mit einer durchschnittlichen Abweichung von 0,3 s für die Nachhallzeit und 0,2 s für die Anfangsnachhallzeit geschätzt. Die vorgeschlagenen Qualitätsparameter wurden anhand einer Klaviertranskription validiert. Mithilfe von simulativ erzeugten Klaviersignalen unterschiedlicher Qualität wurde der Zusammenhang zwischen Transkriptionsgüte und den geschätzten Parametern hergeleitet. Damit ist durch die unabhängigen Qualitätsparameter eine Abschätzung der Transkriptionsgüte sowie eine Einschätzung potenzieller Probleme für die Aufnahmequalität möglich.

Zur zeitabhängigen Instrumentendetektion wurden in dieser Arbeit tiefe KNN-Strukturen mit *Residual*-Modulen und zweidimensionalen Faltungsschichten entwickelt. Diese werden mit den Beträgen der STFT oder CQT des Musiksignals gespeist. Aus den Untersuchungen mit ver-

schiedenen Matrixdimensionen geht hervor, dass eine hohe Frequenzauflösung der Zeit-Frequenz-Darstellungen die Detektionsergebnisse deutlich verbessert, wohingegen eine hohe Zeitauflösung nur zu geringen Verbesserungen führt. Das beste F-Maß von 88,82 % erzielt das Modell mit CQT-Einspeisung der Dimension $(400, 240)$, wodurch es die Ergebnisse vergleichbarer Literaturansätze übertrifft. Gleichzeitig ist die Zeitauflösung der Schätzung um rund 15 % feiner als in der Literatur. Des Weiteren wurden Modelle mit direkter Einspeisung des Zeitsignals analysiert. Deren Ergebnisse sind in F-Maß, *Precision* und *Recall* um mindestens 2 % schlechter als die Modelle mit STFT- und CQT-Einspeisung. Obwohl die Instrumente im Datensatz unterschiedlich oft vorkommen, können die Gesamtergebnisse durch alternative Gütefunktionen wie der gewichteten BCE oder dem *Focal Loss* nicht verbessert werden. Um die Detektionsgüte mit STFT-Einspeisung durch die Einbeziehung von Phaseninformation weiter zu steigern, wurde die zusätzliche Einspeisung der modifizierten Gruppenlaufzeit, des Produktspektrums und der Frequenzfehlermatrix analysiert. Bei einer geringen Frequenzauflösung verbessern alle drei Varianten die Ergebnisse, bei ausreichend guter Frequenzauflösung führen sie aber zu keiner Verbesserung.

Die Separation polyphoner Ensemble-Aufnahmen schätzt die Zeitsignale der zu trennenden Instrumente in dieser Arbeit direkt und verwendet keine Signaltransformation. Es wurden drei unterschiedliche Modellarchitekturen untersucht, welche die Separation in einem Multi-Task-Ansatz, einem gemeinsamen Modell oder unabhängigen Einzelmodellen für jedes Instrument durchführen. Der Multi-Task-Ansatz war den anderen beiden Ansätzen unterlegen. Um die Separationsqualität zu verbessern, wurde in dieser Arbeit die Einspeisung von Zusatzinformation über die zeitabhängigen Aktivitäten aller zu trennenden Instrumente vorgeschlagen. Diese kann vor jedem Block des Encoders erfolgen, weshalb die Separation für alle Einspeiseorte untersucht wurde. Für ein gemeinsames Modell mit 13 Instrumenten beträgt die Verbesserung des SI-SDR durch die Zusatzinformation in allen Einspeiseorten mindestens 1,21 dB, im Falle eines Beispieltrios aus Violine, Trompete und Flöte sogar mindestens 1,44 dB. Die beste Separation erzielt in beiden Fällen die Einspeisung vor Encoderblock 4, welche die Zusatzinformation erst spät im Modell integriert. Bei der Umsetzung mit unabhängigen Einzelmodellen können

vergleichbare SI-SDR-Werte erzielt werden, wenn die Zusatzeinspeisung vor dem GRU-Modul erfolgt. Frühe Einspeisepositionen vor Encoderblock 1 oder 2 führen bei der Separation mit Einzelmodellen sogar zu schlechteren Ergebnissen als ohne Zusatzinformation, sodass sie ungeeignet sind. Damit der Einfluss von fehlerhafter Zusatzinformation abgeschätzt werden kann, wurden mithilfe von zufälligen Invertierungen Fehler simuliert. Für diese fehlerhafte Zusatzinformation führen späte Einspeiseorte, wie vor Encoderblock 3 oder 4, zu deutlich robusteren Separationsergebnissen als frühe Einspeiseorte. Auch wenn beide Architekturvarianten die Überlegenheit einer späteren Einspeisung zeigen, zeichnen sich die gemeinsamen Modelle gegenüber den Einzelmodellen durch eine größere Robustheit aus. Die simulierten Ergebnisse wurden darüber hinaus anhand von echten Schätzungen eines in dieser Arbeit entwickelten Modells zur zeitabhängigen Instrumentendetektion verifiziert. Bei geeigneter Wahl der Schwellenwerte zur Detektion sind die Separationsergebnisse auch mit realer Zusatzinformation besser als ohne Zusatzeinspeisung. Der simulierte Robustheitsverlauf kann nur für die unabhängigen Einzelmodelle validiert werden.

6.2 Ausblick

In dieser Arbeit wurden bewusst monaurale Musikaufnahmen betrachtet, um eine breite Anwendbarkeit zu garantieren. Durch die Hinzunahme von weiteren Aufnahmekanälen könnten Richtungsinformationen miteinbezogen werden, was die MIR-Ansätze vermutlich verbessern würde. Vor allem der weit verbreitete Fall von Stereo-Aufnahmen wäre für alle drei untersuchten Ansätze interessant. Des Weiteren ist im Bereich von maschinellem Lernen die Übertragbarkeit der Ergebnisse auf Eingangsdaten mit anderen Eigenschaften als im Trainingsdatensatz stets eine Herausforderung. Diese sogenannte Generalisierung wurde für die vorgestellten KNN-Modelle noch nicht durchgeführt, da der Fokus dieser Arbeit auf der Entwicklung von verbesserten Verfahren lag. Darüber hinaus werden zur Untersuchung der Generalisierung mehrere geeignete Datensätze benötigt, die teilweise neu zusammengestellt oder sogar neu aufgenommen werden müssten. Deshalb bleibt sie ein Thema für zukünftige Forschungsarbeiten.

Daneben gibt es in den drei behandelten Teilgebieten ebenfalls Ansatzpunkte für weitere Arbeiten. Die Bewertung der Aufnahmequalität wurde hauptsächlich auf Basis von polyphonen Klaviersignalen entwickelt, für die in der Literatur große Datensätze verfügbar sind. Klaviersignale weisen zudem charakteristische Eigenschaften auf, sodass die Übertragung auf mehrere andere Instrumente mit unterschiedlichen Eigenschaften eine mögliche Weiterentwicklung wäre. Dabei ist die Verfügbarkeit von nahezu störungsfreien und hochqualitativen Aufnahmen der Instrumente essenziell, damit das SNR und die Raumakustik-Parameter möglichst exakt vorgegeben werden können. Bei der Instrumentendetektion könnte die parallele Integration der Phasendarstellung im Modell zunächst in einem getrennten Strang erfolgen, damit die zugehörigen Merkmale erst separat extrahiert werden. Da der Ansatz eine typische Klassifikation verfolgt, könnte er darüber hinaus durch halbüberwachtes Lernen mit einer deutlich größeren Datenmenge trainiert werden, für die nur ein Teil der Instrumentenaktivitäten bekannt sind. Des Weiteren könnte das bisher rein auf Zeitsignalen basierende Separationsmodell zu einer hybriden Modellarchitektur erweitert werden, indem seine Ergebnisse mit denen eines zweiten Modells fusioniert werden, welches die Separation auf Basis von Zeit-Frequenz-Darstellungen durchführt. Diese Architektur hat in der Literatur zu besseren Separationsergebnissen geführt, benötigt allerdings auch mehr Ressourcen. Darüber hinaus könnte die Zusatzinformation über die Instrumentenaktivität nicht nur binäre Werte enthalten, sondern in Passagen mit mehreren Instrumenten des gleichen Typs die Anzahl dieser Instrumente repräsentieren.

Anhang

A Detaillierte Ergebnisse zur zeitabhängigen Instrumentendetektion

Bei der Analyse aller Modelle zur zeitabhängigen Instrumentendetektion mit Einspeisung von Zeit-Frequenz-Darstellungen sind in Abschnitt 4.4 teilweise nur die durchschnittlichen F-Maße pro Modell und Matrixdimension angegeben. Die instrumentenspezifischen F-Maße der beiden Modellvarianten ID_ZF_Res und ID_ZF_Deep sowie aller betrachteten Einspeisedimensionen (K, M) der Betrags- und Phasendarstellungen sind in den Tabellen A.1 und A.2 zusammengestellt.

Tabelle A.1 F-Maße des Modells ID_ZF_Res bei Einspeisung von Zeit-Frequenz-Darstellungen mit unterschiedlichen Dimensionen (K, M) und zusätzlichen Phasendarstellungen.

	(K, M)	Piano	Violine	Viola	Cello	Klarin.	Fagott	Horn	Ø
STFT	(513, 240)	98,22 %	95,51 %	78,81 %	91,36 %	87,75 %	81,52 %	77,47 %	87,23 %
	(2049, 60)	98,08 %	95,09 %	**83,11 %**	91,57 %	**90,64 %**	**84,02 %**	76,19 %	88,39 %
	(2049, 240)	98,11 %	95,08 %	81,05 %	91,73 %	90,09 %	83,58 %	79,48 %	88,44 %
STFT + ModGD	(513, 240)	98,04 %	95,66 %	80,76 %	91,59 %	88,43 %	82,79 %	77,86 %	87,88 %
	(2049, 60)	98,17 %	95,55 %	81,54 %	90,84 %	**90,64 %**	82,58 %	78,74 %	88,30 %
	(2049, 240)	98,09 %	95,62 %	81,41 %	91,15 %	90,24 %	82,71 %	78,88 %	88,30 %
STFT + PS	(513, 240)	97,97 %	95,65 %	83,03 %	92,15 %	88,09 %	82,64 %	78,60 %	88,30 %
	(2049, 60)	98,22 %	95,23 %	81,60 %	91,80 %	89,47 %	83,00 %	80,87 %	88,60 %
	(2049, 240)	97,95 %	95,31 %	79,88 %	91,07 %	89,75 %	83,21 %	81,89 %	88,44 %
STFT + Ψ	(513, 240)	**98,33 %**	**95,89 %**	82,40 %	91,36 %	88,14 %	81,82 %	78,93 %	88,12 %
	(2049, 60)	98,01 %	94,75 %	79,19 %	91,96 %	90,13 %	83,79 %	78,98 %	88,11 %
	(2049, 240)	97,99 %	95,19 %	78,82 %	91,62 %	89,75 %	83,00 %	79,93 %	88,04 %
CQT	(88, 60)	97,71 %	95,39 %	79,80 %	91,19 %	86,44 %	81,74 %	76,07 %	86,91 %
	(88, 240)	97,65 %	94,03 %	80,01 %	91,40 %	87,94 %	82,76 %	76,37 %	87,17 %
	(400, 60)	97,66 %	95,64 %	81,38 %	**92,49 %**	86,45 %	81,93 %	**82,10 %**	88,24 %
	(400, 240)	98,31 %	95,73 %	81,72 %	92,27 %	89,27 %	83,54 %	80,93 %	**88,82 %**

Tabelle A.2 F-Maße des Modells ID_ZF_Deep bei Einspeisung von Zeit-Frequenz-Darstellungen mit unterschiedlichen Dimensionen (K, M) und zusätzlichen Phasendarstellungen.

	(K, M)	Piano	Violine	Viola	Cello	Klarin.	Fagott	Horn	Ø
STFT	(513, 240)	**98,35 %**	95,50 %	82,29 %	91,10 %	89,02 %	81,58 %	74,32 %	87,45 %
	(2049, 60)	98,19 %	95,48 %	82,67 %	90,94 %	88,85 %	82,45 %	77,79 %	88,05 %
	(2049, 240)	98,27 %	94,80 %	82,32 %	91,81 %	87,90 %	82,35 %	73,12 %	87,22 %
STFT + ModGD	(513, 240)	98,29 %	95,47 %	77,33 %	91,86 %	88,57 %	82,36 %	74,83 %	86,96 %
	(2049, 60)	97,89 %	95,08 %	80,35 %	89,37 %	90,01 %	83,76 %	78,85 %	87,90 %
	(2049, 240)	98,15 %	94,82 %	79,29 %	90,45 %	88,93 %	84,25 %	79,54 %	87,92 %
STFT + PS	(513, 240)	97,94 %	95,34 %	82,65 %	91,27 %	87,89 %	82,94 %	80,62 %	88,38 %
	(2049, 60)	97,88 %	94,94 %	76,83 %	91,47 %	88,80 %	82,75 %	77,88 %	87,22 %
	(2049, 240)	97,91 %	95,53 %	77,83 %	91,70 %	89,15 %	83,46 %	**80,73 %**	88,04 %
STFT + Ψ	(513, 240)	98,09 %	95,25 %	79,38 %	90,88 %	87,53 %	82,42 %	77,26 %	87,26 %
	(2049, 60)	**98,35 %**	95,31 %	**83,03 %**	91,98 %	**90,59 %**	82,29 %	79,65 %	**88,74 %**
	(2049, 240)	98,23 %	**95,99 %**	82,16 %	91,93 %	88,87 %	81,97 %	78,99 %	88,31 %
CQT	(88, 60)	97,30 %	93,98 %	79,49 %	90,81 %	85,23 %	79,16 %	75,58 %	85,94 %
	(88, 240)	97,91 %	95,57 %	81,05 %	91,77 %	89,45 %	82,45 %	78,19 %	88,05 %
	(400, 60)	97,99 %	94,30 %	80,59 %	91,39 %	88,42 %	**84,88 %**	78,76 %	88,05 %
	(400, 240)	97,79 %	94,96 %	81,09 %	**92,44 %**	87,00 %	83,18 %	79,55 %	88,00 %

B Detaillierte Ergebnisse der Separationsmodelle

Zur Separation von Ensemble-Aufnahmen unterschiedlicher Musikinstrumente werden in Abschnitt 5.5.1 gemeinsame Separationsmodelle mit 13 und 3 Instrumenten untersucht. Durch die zusätzliche Einspeisung der zeitabhängigen Aktivität aller betrachteten Instrumente kann die Separationsqualität gesteigert werden. Die Bewertung der separierten Instrumentensignale erfolgt in Abschnitt 5.5.1 anhand des SI-SDR der sechs im URMP-Testdatensatz vorhandenen Instrumente. In Tabelle B.1 sind die SI-SDR-Werte aller analysierten Modellvarianten mit 13 Instrumenten und zeitabhängiger, statischer sowie ohne Zusatzinformation zusammengestellt. Um eine Einschätzung der Verbesserung durch den augmentierten URMP-Datensatz geben zu können, enthält die letzte Spalte die Verbesserung des durchschnittlichen SI-SDR-Werts gegenüber dem Ergebnis des mit dem nicht augmentierten URMP-Datensatz trainierten Modells. Eine detailliertere Bewertung der Separationsqualität ermöglichen die beiden Metriken SI-SIR und SI-SAR, welche für die gemeinsamen Separationsmodelle mit 13 Instrumenten aus Tabelle B.1 in Tabelle B.2 und Tabelle B.3 aufgelistet sind.

Die Separationsmodelle des vorab definierten Beispieltrios aus Violine, Trompete und Flöte werden anhand dieser drei Instrumente im URMP-Testdatensatz evaluiert. Alle Modellvarianten werden in Abschnitt 5.5.1 mithilfe des SI-SDR bewertet, dessen Werte in Tabelle B.4 zusammengestellt sind. Auch hier ist jeweils die Verbesserung durch die Nutzung des augmentierten URMP-Datensatzes im Training mit angegeben. Darüber hinaus sind die SI-SIR-Werte aller Modelle des Trios in Tabelle B.5 aufgelistet. Dabei werden für das SI-SIR der Flöte nur die ersten beiden der drei Teststücke berücksichtigt, da im dritten Stück keine Trompete und Violine vorkommen, sondern nur andere Instrumente und die Flöte. Für das SI-SIR des Trios sind die anderen Instrumente aber irrelevant, weshalb

die Trennung „perfekt“ funktioniert und für das dritte Teststück folglich sehr hohe, unrealistische SI-SIR-Werte berechnet werden. Tabelle B.6 enthält mit den SI-SAR-Werten die dritte Metrik der Separationsmodelle für das Trio, die wieder für alle Teststücke errechnet wird.

Als Alternative zur gemeinsamen Modellarchitektur werden in Abschnitt 5.5.2 Separationssysteme aus unabhängigen Einzelmodellen jedes zu trennenden Instruments untersucht. Auch hier kann die Separationsqualität durch die Integration von zeitabhängiger Instrumentenaktivität verbessert werden. Die Ergebnisse aller untersuchten Modellvarianten sowie des Literaturansatzes Open-Unmix [132] sind mithilfe der SI-SDR-Werte in Tabelle B.7 aufgeführt. Durch die Verwendung des augmentierten URMP-Datensatzes im Training kann das durchschnittliche Separationsergebnis gegenüber dem Fall des nicht augmentierten Datensatzes gesteigert werden. Die Verbesserung ist jeweils in der letzten Spalte angegeben. Eine genauere Analyse der Separationsergebnisse ermöglichen die SI-SIR-Werte aller Einzelmodelle in Tabelle B.8 sowie die zugehörigen SI-SAR-Werte, welche in Tabelle B.9 aufgelistet sind.

Auffällig ist, dass alle SI-SAR-Werte für den Fall ohne Separation (‚Mix‘) extrem hoch sind. Da keine Separation durchgeführt wird, entspricht das Eingangssignal dem Ausgangssignal und es treten keine Artefakte auf. Das SI-SAR ist nach Gleichung 5.3 aber ein Maß für das Verhältnis zwischen den Signalenergien des zu trennenden Zielsignals und der im geschätzten Signal vorhandenen Artefakte. Folglich detektiert der Algorithmus zur Berechnung des SI-SAR korrekterweise sehr wenig Artefakte, woraus die sehr hohen SI-SAR-Werte resultieren.

Tabelle B.1 SI-SDR (in dB) aller Instrumente des URMP-Testdatensatzes für gemeinsame Modelle mit 13 Instrumenten und unterschiedlichen Einspeisungen von Zusatzinformation.

Zusatz-information	Einspeise-position	Fagott	Flöte	Oboe	Sax.	Tromp.	Violine	Ø	Verbess. Augm.
ohne	-	–7,60	3,06	–0,99	–6,62	–0,28	–1,07	–2,25	+3,15
zeitabhängig	Enc. 1	**1,31**	1,89	–0,75	–5,86	**3,95**	**1,62**	0,36	+2,74
	Enc. 2	0,26	0,82	–2,28	–3,80	3,80	–1,51	–0,45	+8,17
	Enc. 3	–2,77	3,47	–3,04	–3,80	1,60	–1,68	–1,04	+2,45
	Enc. 4	0,86	**4,74**	–0,42	**–3,53**	1,36	0,65	**0,61**	+1,98
	GRU	0,62	4,01	–1,38	–3,80	1,00	–0,16	0,05	+4,04
statisch	Enc. 1	0,69	1,07	–0,93	–7,13	3,28	0,33	–0,45	+2,53
	Enc. 2	–0,10	0,20	–2,38	–5,00	3,17	–2,45	–1,09	+7,89
	Enc. 3	–3,08	2,43	–2,97	–5,00	1,08	–2,77	–1,72	+2,36
	Enc. 4	0,94	3,73	**–0,39**	–4,33	0,94	–0,42	0,08	+1,93
	GRU	0,01	3,01	–1,56	–4,48	0,45	–1,22	–0,63	+4,58
Mix	-	–3,51	1,29	–11,29	–6,17	–4,12	–4,98	–4,80	-
Multipl.	Ausgang	–7,60	3,17	–0,96	–6,59	–0,18	–1,12	–2,22	+3,13

Tabelle B.2 SI-SIR (in dB) aller Instrumente des URMP-Testdatensatzes für gemeinsame Modelle mit 13 Instrumenten und unterschiedlichen Einspeisungen von Zusatzinformation.

Zusatz-information	Einspeise-position	Fagott	Flöte	Oboe	Sax.	Tromp.	Violine	Ø	Verbess. Augm.
ohne	-	2,23	13,72	11,50	–0,53	9,69	7,92	7,42	+1,66
zeitabhängig	Enc. 1	**8,92**	15,59	5,93	–1,76	12,57	7,08	8,06	+4,72
	Enc. 2	6,04	15,60	3,19	–1,00	12,35	**8,17**	7,39	+7,53
	Enc. 3	5,57	16,12	2,10	–0,69	**15,11**	7,38	7,60	+4,18
	Enc. 4	7,28	**16,21**	8,29	–1,63	13,06	7,67	**8,48**	+4,07
	GRU	5,76	14,78	5,94	**0,88**	12,26	6,45	7,68	+0,70
statisch	Enc. 1	8,15	12,89	5,78	–3,58	11,09	5,62	6,66	+4,56
	Enc. 2	5,39	12,57	3,13	–2,41	10,62	6,79	6,01	+7,10
	Enc. 3	5,21	12,83	2,18	–2,21	12,98	5,85	6,14	+3,95
	Enc. 4	6,74	13,08	8,75	–2,65	11,36	6,47	7,29	+4,00
	GRU	5,39	12,39	5,82	0,21	10,42	5,26	6,58	+2,05
Mix	-	–3,51	1,29	–11,29	–6,17	–4,12	–4,98	–4,80	-
Multipl.	Ausgang	2,28	15,11	**11,64**	–0,44	10,50	8,02	7,85	+1,69

Tabelle B.3 SI-SAR (in dB) aller Instrumente des URMP-Testdatensatzes für gemeinsame Modelle mit 13 Instrumenten und unterschiedlichen Einspeisungen von Zusatzinformation.

Zusatz-information	Einspeise-position	Fagott	Flöte	Oboe	Sax.	Tromp.	Violine	Ø	Verbess. Augm.
ohne	-	–7,13	3,52	–0,74	–5,40	0,18	–0,45	–1,67	+3,19
zeitabhängig	Enc. 1	2,13	2,13	**0,30**	–3,71	**4,59**	**3,10**	1,42	+2,18
	Enc. 2	1,59	1,01	–0,84	–0,56	4,45	–1,02	0,77	+7,65
	Enc. 3	–2,09	3,76	–1,45	–0,88	1,79	–1,10	0,01	+2,32
	Enc. 4	1,99	**5,07**	0,21	**0,99**	1,66	1,61	**1,92**	+1,79
	GRU	2,21	4,44	–0,49	–1,99	1,33	0,91	1,07	+3,88
statisch	Enc. 1	1,55	1,42	0,11	–4,60	4,06	1,87	0,74	+1,80
	Enc. 2	1,35	0,50	–0,95	–1,52	4,03	–1,90	0,25	+7,03
	Enc. 3	–2,38	2,90	–1,39	–1,75	1,37	–2,12	–0,56	+2,11
	Enc. 4	**2,27**	4,27	0,17	0,61	1,35	0,57	1,54	+1,59
	GRU	1,50	3,58	–0,68	–2,68	0,92	–0,11	0,42	+4,26
Mix	-	**149,36**	**144,42**	**142,87**	**147,36**	**139,20**	**137,36**	**143,43**	-
Multipl.	Ausgang	–7,12	3,51	–0,72	–5,39	0,21	–0,53	–1,67	+3,19

Tabelle B.4 SI-SDR (in dB) der Instrumente des definierten Beispieltrios für gemeinsame Modelle dieses Trios und unterschiedlichen Einspeisungen von Zusatzinformation.

Zusatz-information	Einspeise-position	Flöte	Tromp.	Violine	∅	Verbess. Augm.
ohne	-	2,60	–0,26	–0,07	0,76	+3,97
zeitabhängig	Enc. 1	**5,06**	0,77	1,69	2,51	+5,36
	Enc. 2	4,00	1,20	1,40	2,20	+5,19
	Enc. 3	4,78	1,23	2,00	2,67	+4,86
	Enc. 4	4,96	**1,46**	**2,38**	**2,93**	+4,42
	GRU	4,80	0,98	2,35	2,71	+4,95
statisch	Enc. 1	3,68	0,40	0,20	1,43	+4,84
	Enc. 2	3,25	0,94	0,37	1,52	+5,63
	Enc. 3	3,85	0,94	1,08	1,96	+5,08
	Enc. 4	4,14	1,16	1,41	2,24	+4,29
	GRU	3,79	0,42	0,90	1,70	+4,74
Mix	-	1,29	–4,12	–4,98	–2,60	-
Multipl.	Ausgang	2,70	–0,12	–0,12	0,82	+3,91

Tabelle B.5 SI-SIR (in dB) der Instrumente des definierten Beispieltrios für gemeinsame Modelle dieses Trios und unterschiedlichen Einspeisungen von Zusatzinformation.

Zusatz-information	Einspeise-position	Flöte	Tromp.	Violine	∅	Verbess. Augm.
ohne	-	14,33	9,38	7,84	10,52	+2,58
zeitabhängig	Enc. 1	16,63	11,52	8,26	12,14	+5,31
	Enc. 2	**17,48**	12,46	7,86	12,60	+6,27
	Enc. 3	17,43	11,83	8,59	12,62	+8,01
	Enc. 4	17,37	**12,78**	**9,49**	**13,21**	+8,00
	GRU	17,23	12,14	9,38	12,92	+4,33
statisch	Enc. 1	13,64	10,10	5,92	9,89	+5,44
	Enc. 2	13,97	11,20	6,76	10,64	+6,52
	Enc. 3	14,29	10,55	7,42	10,75	+8,07
	Enc. 4	14,61	11,36	8,42	11,46	+7,94
	GRU	13,69	10,33	7,67	10,56	+4,94
Mix	-	2,40	–4,12	–4,98	–2,23	-
Multipl.	Ausgang	15,68	10,36	7,91	11,32	+2,52

Tabelle B.6 SI-SAR (in dB) der Instrumente des definierten Beispieltrios für gemeinsame Modelle dieses Trios und unterschiedlichen Einspeisungen von Zusatzinformation.

Zusatz-information	Einspeise-position	Flöte	Tromp.	Violine	∅	Verbess. Augm.
ohne	-	2,81	0,24	0,71	1,25	+4,11
zeitabhängig	Enc. 1	**5,30**	1,15	2,78	3,08	+5,42
	Enc. 2	4,16	1,54	2,52	2,74	+5,00
	Enc. 3	4,97	1,63	3,07	3,22	+4,47
	Enc. 4	5,17	**1,79**	3,32	**3,43**	+3,83
	GRU	5,02	1,32	**3,34**	3,23	+4,93
statisch	Enc. 1	3,99	0,89	1,55	2,14	+4,76
	Enc. 2	3,54	1,37	1,50	2,14	+5,38
	Enc. 3	4,16	1,45	2,23	2,61	+4,59
	Enc. 4	4,47	1,60	2,38	2,82	+3,52
	GRU	4,13	0,88	1,95	2,32	+4,65
Mix	-	**93,78**	**139,20**	**137,36**	**123,45**	-
Multipl.	Ausgang	2,85	0,29	0,63	1,26	+4,07

Tabelle B.7 SI-SDR (in dB) aller Instrumente des URMP-Testdatensatzes für unabhängige Modelle der 13 Instrumente mit unterschiedlichen Einspeisungen von Zusatzinformation.

Zusatz-information	Einspeise-position	Fagott	Flöte	Oboe	Sax.	Tromp.	Violine	Ø	Verbess. Augm.
ohne	-	–44,81	1,99	–2,28	–14,81	3,38	–1,00	–9,59	+0,56
zeitabhängig	Enc. 1	–5,47	0,46	–0,98	–6,21	1,58	–2,36	–2,16	+2,00
	Enc. 2	–6,69	1,63	–0,72	–3,98	1,22	0,10	–1,41	+1,00
	Enc. 3	–7,05	2,42	**–0,35**	**–3,39**	4,03	0,42	–0,65	+1,35
	Enc. 4	–6,63	1,90	–0,98	–4,76	4,21	0,69	–0,93	+1,00
	GRU	–4,00	**3,86**	–1,32	–4,78	**4,38**	0,68	**–0,20**	+1,15
statisch	Enc. 1	–5,58	–0,40	–1,25	–7,19	0,85	–3,34	–2,82	+1,59
	Enc. 2	–6,80	0,65	–0,81	–4,76	0,79	–0,62	–1,93	+0,84
	Enc. 3	–7,13	1,37	–0,49	–4,48	3,21	–0,08	–1,27	+1,12
	Enc. 4	–6,70	1,19	–1,10	–5,31	3,37	0,77	–1,30	+1,02
	GRU	–4,05	2,55	–1,40	–5,29	3,70	0,70	–0,63	+1,05
Mix	-	**–3,51**	1,29	–11,29	–6,17	–4,12	–4,98	–4,80	-
Multipl.	Ausgang	–44,54	2,12	–2,28	–14,77	3,46	–1,05	–9,51	+0,53
Op.-U. [132]	-	–7,52	1,92	–0,45	–30,92	0,08	**2,53**	–5,73	-

Tabelle B.8 SI-SIR (in dB) aller Instrumente des URMP-Testdatensatzes für unabhängige Modelle der 13 Instrumente mit unterschiedlichen Einspeisungen von Zusatzinformation.

Zusatz-information	Einspeise-position	Fagott	Flöte	Oboe	Sax.	Tromp.	Violine	Ø	Verbess. Augm.
ohne	-	–7,37	14,80	15,77	0,96	13,83	7,46	7,58	+1,16
zeitabhängig	Enc. 1	**7,60**	15,05	13,05	–1,69	**14,94**	6,07	9,17	+4,50
	Enc. 2	5,07	16,10	12,85	–0,26	13,91	7,26	9,15	+3,33
	Enc. 3	3,75	15,01	14,06	0,61	14,48	8,88	9,47	+3,46
	Enc. 4	5,72	13,88	13,37	0,64	14,88	9,08	**9,60**	+3,93
	GRU	7,20	16,14	9,91	–0,49	14,57	8,79	9,35	+3,11
statisch	Enc. 1	7,33	9,20	12,35	–3,34	12,75	5,10	7,23	+3,58
	Enc. 2	4,75	10,73	12,42	–1,61	11,87	6,65	7,47	+2,88
	Enc. 3	3,36	10,13	13,59	–1,13	12,50	8,43	7,82	+3,09
	Enc. 4	5,33	9,54	12,92	–0,66	12,79	9,52	8,24	+3,79
	GRU	6,86	11,19	9,59	–1,48	12,78	8,86	7,97	+2,98
Mix	-	–3,51	1,29	–11,29	–6,17	–4,12	–4,98	–4,80	-
Multipl. Op.-U. [132]	Ausgang	–7,59	**16,30**	**15,80**	**1,34**	14,56	7,50	7,99	+1,09
	-	4,51	15,06	14,65	–17,06	5,60	**12,99**	5,96	-

Tabelle B.9 SI-SAR (in dB) aller Instrumente des URMP-Testdatensatzes für unabhängige Modelle der 13 Instrumente mit unterschiedlichen Einspeisungen von Zusatzinformation.

Zusatz-information	Einspeise-position	Fagott	Flöte	Oboe	Sax.	Tromp.	Violine	Ø	Verbess. Augm.
ohne	-	–44,81	2,29	–2,22	–14,69	3,79	–0,33	–9,33	+0,46
zeitabhängig	Enc. 1	–5,25	0,66	–0,80	–4,32	1,79	–1,67	–1,60	+1,74
	Enc. 2	–6,39	1,82	–0,53	–1,57	1,46	1,03	–0,70	+0,65
	Enc. 3	–6,67	2,71	**–0,19**	**–1,18**	4,44	1,09	0,03	+0,92
	Enc. 4	–6,37	2,22	–0,81	–3,29	4,59	1,37	–0,38	+0,19
	GRU	**–3,65**	**4,15**	–0,98	–2,76	**4,81**	1,42	**0,50**	+0,42
statisch	Enc. 1	–5,35	0,11	–1,06	–4,88	1,14	–2,66	–2,12	+1,29
	Enc. 2	–6,48	1,10	–0,60	–1,89	1,14	0,28	–1,07	+0,40
	Enc. 3	–6,72	2,00	–0,32	–1,79	3,75	0,58	–0,42	+0,57
	Enc. 4	–6,42	1,87	–0,92	–3,49	3,90	1,40	–0,61	+0,04
	GRU	–3,69	3,21	–1,04	–2,94	4,27	1,42	0,20	+0,05
Mix	-	**149,36**	**144,42**	**142,87**	**147,36**	**139,20**	**137,36**	**143,43**	-
Multipl.	Ausgang	–44,54	2,33	–2,21	–14,66	3,82	–0,40	–9,28	+0,45
Op.-U. [132]	-	–7,24	2,17	–0,32	–30,74	1,52	**2,96**	–5,28	-

Literaturverzeichnis

[1] **C. C. Aggarwal**. *Neural networks and deep learning*. Springer Cham, 2018.

[2] **H. Banno, J. Lu, S. Nakamura, K. Shikano und H. Kawahara**. *Efficient representation of short-time phase based on group delay*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Bd. 2. 1998, S. 841–864.

[3] **D. Barry, E. Coyle, D. Fitzgerald und R. Lawlor**. *Single channel source separation using short-time independent component analysis*. In: *Audio Engineering Society Convention 119*. 2005, 6603.

[4] **R. M. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam und J. P. Bello**. *MedleyDB: A multitrack dataset for annotation-intensive MIR research*. In: *15th International Society for Music Information Retrieval Conference*. 2014, S. 155–160.

[5] **J. J. Bosch, J. Janer, F. Fuhrmann und P. Herrera**. *A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals*. In: *13th International Society for Music Information Retrieval Conference*. 2012, S. 559–564.

[6] **J. C. Brown**. *Calculation of a constant Q spectral transform*. In: *Journal of the Acoustical Society of America* 89.1 (1991), S. 425–434.

[7] **J. C. Brown und M. S. Puckette**. *An efficient algorithm for the calculation of a constant Q transform*. In: *Journal of the Acoustical Society of America* 92.5 (1992), S. 2698–2701.

[8] **O. Calin**. *Deep learning architectures: A mathematical approach*. Springer Cham, 2020.

[9] **G. Cantisani, A. Ozerov, S. Essid und G. Richard**. *User-guided one-shot deep model adaptation for music source separation*. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2021, S. 111–115.

[10] **P. Chandna, M. Miron, J. Janer und E. Gómez**. *Monoaural audio source separation using deep convolutional neural networks*. In: *Latent Variable Analysis and Signal Separation*. Hrsg. von **P. Tichavský, M. Babaie-Zadeh, O. J. J. Michel und N. Thirion-Moreau**. Springer Cham, 2017, S. 258–266.

[11] **K. W. Cheuk, K. Agres und D. Herremans**. *The impact of audio input representations on neural network based music transcription*. In: *International Joint Conference on Neural Networks*. 2020, S. 1–6.

[12] **K. Choi, G. Fazekas, K. Cho und M. Sandler**. *A tutorial on deep learning for music information retrieval*. In: *arXiv preprint arXiv:1709.04396* (2017).

[13] **W. Choi, M. Kim, J. Chung und S. Jung**. *LaSAFT: Latent source attentive frequency transformation for conditioned source separation*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2021, S. 171–175.

[14] **J. Chung, C. Gulcehre, K. Cho und Y. Bengio**. *Empirical evaluation of gated recurrent neural networks on sequence modeling*. In: *arXiv preprint arXiv:1412.3555* (2014).

[15] **N. B. H. Croghan, K. H. Arehart und J. M. Kates**. *Quality and loudness judgments for music subjected to compression limiting*. In: *The Journal of the Acoustical Society of America* 132.2 (2012), S. 1177–1188.

[16] **Y. N. Dauphin, A. Fan, M. Auli und D. Grangier**. *Language modeling with gated convolutional networks*. In: *34th International Conference on Machine Learning*. 2017, S. 933–941.

[17] **A. Défossez**. *Hybrid spectrogram and waveform source separation*. In: *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*. 2021.

[18] **A. Défossez, N. Usunier, L. Bottou und F. Bach**. *Music source separation in the waveform domain*. In: *arXiv preprint arXiv:1911.13254* (2019).

[19] **L. Deng**. *Artificial intelligence in the rising wave of deep learning: The historical path and future outlook [perspectives]*. In: *IEEE Signal Processing Magazine* 35.1 (2018), S. 180–177.

[20] **R. Deutsch und L. J. Deutsch**. *ADSR envelope generator*. In: *The Journal of the Acoustical Society of America* 66.3 (1979), S. 936.

[21] **S. Dieleman und B. Schrauwen**. *End-to-end learning for music audio*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2014, S. 6964–6968.

[22] **S. Diether, L. Bruderer, A. Streich und H.-A. Loeliger**. *Efficient blind estimation of subband reverberation time from speech in non-diffuse environments*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2015, S. 743–747.

[23] **A. Diment, T. Heittola und T. Virtanen**. *Semi-supervised learning for musical instrument recognition*. In: *21st European Signal Processing Conference*. 2013, S. 1–5.

[24] **A. Diment, P. Rajan, T. Heittola und T. Virtanen**. *Modified group delay feature for musical instrument recognition*. In: *10th International Symposium on Computer Music Multidisciplinary Research*. 2013, S. 431–438.

[25] **S. R. Dubey, S. K. Singh und B. B. Chaudhuri**. *Activation functions in deep learning: A comprehensive survey and benchmark*. In: *Neurocomputing* 503 (2022), S. 92–108.

[26] **J. Eaton, N. D. Gaubitch und P. A. Naylor**. *Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, S. 161–165.

[27] **V. Emiya, R. Badeau und B. David**. *Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle*. In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.6 (2009), S. 1643–1654.

[28] **V. Emiya, E. Vincent, N. Harlander und V. Hohmann**. *Subjective and objective quality assessment of audio source separation*. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), S. 2046–2057.

[29] **W. Ertel und N. T. Black**. *Grundkurs künstliche Intelligenz: Eine praxisorientierte Einführung*. Springer Vieweg Wiesbaden, 2021.

[30] **S. Essid, G. Richard und B. David**. *Musical instrument recognition on solo performances*. In: *12th European Signal Processing Conference*. 2004, S. 1289–1292.

[31] **S. Essid, G. Richard und B. David**. *Instrument recognition in polyphonic music based on automatic taxonomies*. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.1 (2006), S. 68–80.

[32] **C. Févotte, E. Vincent und A. Ozerov**. *Single-channel audio source separation with NMF: divergences, constraints and algorithms*. In: *Audio Source Separation*. Hrsg. von **S. Makino**. Springer Cham, 2018, S. 1–26.

[33] **H. Fletcher**. *Normal vibration frequencies of a stiff piano string*. In: *The Journal of the Acoustical Society of America* 36.1 (1964), S. 203–209.

[34] **H. Fletcher und W. A. Munson**. *Loudness, its definition, measurement and calculation*. In: *The Bell System Technical Journal* 12.4 (1933), S. 377–430.

[35] **H. Flores Garcia, A. Aguilar, E. Manilow und B. Pardo**. *Leveraging hierarchical structures for few-shot musical instrument recognition*. In: *22nd International Society for Music Information Retrieval Conference*. 2021, S. 220–228.

[36] **X. Glorot, A. Bordes und Y. Bengio**. *Deep sparse rectifier neural networks*. In: *14th International Conference on Artificial Intelligence and Statistics*. 2011, S. 315–323.

[37] **I. Goodfellow, Y. Bengio und A. Courville**. *Deep learning*. MIT press, 2016.

[38] **E. M. Grais und M. D. Plumbley**. *Single channel audio source separation using convolutional denoising autoencoders*. In: *IEEE Global Conference on Signal and Information Processing*. 2017, S. 1265–1269.

[39] **D. W. Griffin und J. S. Lim**. *Signal estimation from modified short-time Fourier transform*. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.2 (1984), S. 236–243.

[40] **S. Gururani, M. Sharma und A. Lerch**. *An attention mechanism for musical instrument recognition*. In: *20th International Society for Music Information Retrieval Conference*. 2019, S. 83–90.

[41] **S. Hamawaki, S. Funasawa, J. Katto, H. Ishizaki, K. Hoashi und Y. Takishima**. *Feature analysis and normalization approach for robust content-based music retrieval to encoded audio with different bit rates*. In: *International Conference on Multimedia Modeling*. 2009, S. 298–309.

[42] **Y. Han, J. Kim und K. Lee**. *Deep convolutional neural networks for predominant instrument recognition in polyphonic music*. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.1 (2017), S. 208–221.

[43] **F. J. Harris**. *On the use of windows for harmonic analysis with the discrete Fourier transform*. In: *Proceedings of the IEEE* 66.1 (1978), S. 51–83.

[44] **C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore und D. Eck**. *Onsets and frames: Dual-objective piano transcription*. In: *19th International Society for Music Information Retrieval Conference*. 2018, S. 50–57.

[45] **K. He, X. Zhang, S. Ren und J. Sun**. *Deep residual learning for image recognition*. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016, S. 770–778.

[46] **R. Hennequin, A. Khlif, F. Voituret und M. Moussallam**. *Spleeter: a fast and efficient music source separation tool with pre-trained models*. In: *Journal of Open Source Software* 5.50, 2154 (2020).

[47] **G. E. Hinton und R. R. Salakhutdinov**. *Reducing the dimensionality of data with neural networks*. In: *Science* 313.5786 (2006), S. 504–507.

[48] **S. Hochreiter und J. Schmidhuber**. *Long short-term memory*. In: *Neural Computation* 9.8 (1997), S. 1735–1780.

[49] **N. Holighaus, M. Dörfler, G. A. Velasco und T. Grill**. *A framework for invertible, real-time constant-Q transforms*. In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.4 (2013), S. 775–785.

[50] **M.-P. Hosseini, S. Lu, K. Kamaraj, A. Slowikowski und H. C. Venkatesh**. *Deep learning architectures*. In: *Deep Learning: Concepts and Architectures*. Hrsg. von **W. Pedrycz und S.-M. Chen**. Springer Cham, 2020, S. 1–24.

[51] **Y.-N. Hung, Y.-A. Chen und Y.-H. Yang**. *Multitask learning for frame-level instrument recognition*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2019, S. 381–385.

[52] **Y.-N. Hung und A. Lerch**. *Multitask learning for instrument activation aware music source separation*. In: *21st International Society for Music Information Retrieval Conference*. 2020, S. 748–755.

[53] **Y.-N. Hung und Y.-H. Yang**. *Frame-level instrument recognition by timbre and pitch*. In: *19th International Society for Music Information Retrieval Conference*. 2018, S. 135–142.

[54] **S. Ioffe und C. Szegedy**. *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. In: *32nd International Conference on Machine Learning*. 2015, S. 448–456.

[55] **G. James, D. Witten, T. Hastie und R. Tibshirani**. *An introduction to statistical learning*. Springer New York, NY, 2013.

[56] **A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar und T. Weyde**. *Singing voice separation with deep U-Net convolutional networks*. In: *18th International Society for Music Information Retrieval Conference*. 2017, S. 745–751.

[57] **D. H. Johnson**. *The signal processing information base: Noise data*. URL: http://spib.linse.ufsc.br/noise.html (besucht am 01. 02. 2023).

[58] **I. Kaminskyj und A. Materka**. *Automatic source identification of monophonic musical instrument sounds*. In: *International Conference on Neural Networks*. Bd. 1. 1995, S. 189–194.

[59] **M. Kawamura, T. Nakamura, D. Kitamura, H. Saruwatari, Y. Takahashi und K. Kondo**. *Differentiable digital signal processing mixture model for synthesis parameter extraction from mixture of harmonic sounds*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2022, S. 941–945.

[60] **A. Kendall, Y. Gal und R. Cipolla**. *Multi-task learning using uncertainty to weigh losses for scene geometry and semantics*. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2018, S. 7482–7491.

[61] **P. Kendrick, T. J. Cox, Y. Zhang, J. A. Chambers und F. F. Li**. *Room acoustic parameter extraction from music signals*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Bd. 5. 2006, S. V801–V804.

[62] **P. Kendrick, T. J. Cox, F. F. Li, Y. Zhang und J. A. Chambers**. *Monaural room acoustic parameters from music and speech*. In: *The Journal of the Acoustical Society of America* 124.1 (2008), S. 278–287.

[63] **U. Kiencke, M. Schwarz und T. Weickert**. *Signalverarbeitung*. Oldenbourg Wissenschaftsverlag, 2008.

[64] **C. Kim und R. M. Stern**. *Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis*. In: *9th Annual Conference of the International Speech Communication Association*. 2008, S. 2598–2601.

[65] **M. Kim, W. Choi, J. Chung, D. Lee und S. Jung**. *KUIELab-MDX-Net: A two-stream neural network for music demixing*. In: *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*. 2021.

[66] **D. P. Kingma und J. Ba**. *Adam: A method for stochastic optimization*. In: *arXiv preprint arXiv:1412.6980* (2014).

[67] **S. Kirkup**. *The boundary element method in acoustics: A survey*. In: *Applied Sciences* 9.8, 1642 (2019).

[68] **T. Kitahara, M. Goto, K. Komatani, T. Ogata und H. G. Okuno**. *Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps*. In: *EURASIP Journal on Advances in Signal Processing* 2007.1 (2006).

[69] **Q. Kong, Y. Cao, H. Liu, K. Choi und Y. Wang**. *Decoupling magnitude and phase estimation with deep ResUNet for music source separation*. In: *22nd International Society for Music Information Retrieval Conference*. 2021, S. 342–349.

[70] **A. Kratimenos, K. Avramidis, C. Garoufis, A. Zlatintsi und P. Maragos**. *Augmentation methods on monophonic audio for instrument classification in polyphonic music*. In: *28th European Signal Processing Conference*. 2020, S. 156–160.

[71] **H. Kuttruff**. *Room acoustics*. CRC Press, 2016.

[72] **J. Le Roux, S. Wisdom, H. Erdogan und J. R. Hershey**. *SDR–half-baked or well done?* In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2019, S. 626–630.

[73] **Y. LeCun, Y. Bengio und G. Hinton**. *Deep learning*. In: *Nature* 521 (2015), S. 436–444.

[74] **R. Lerch, G. Sessler und D. Wolf**. *Technische Akustik: Grundlagen und Anwendungen*. Springer Berlin, Heidelberg, 2009.

[75] **B. Li, X. Liu, K. Dinesh, Z. Duan und G. Sharma**. *Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications*. In: *IEEE Transactions on Multimedia* 21.2 (2019), S. 522–535.

[76] **P. Li, J. Qian und T. Wang**. *Automatic instrument recognition in polyphonic music using convolutional neural networks*. In: *arXiv preprint arXiv:1511.05520* (2015).

[77] **L. Lin, Q. Kong, J. Jiang und G. Xia**. *A unified model for zero-shot music source separation, transcription and synthesis*. In: *22nd International Society for Music Information Retrieval Conference*. 2021, S. 381–388.

[78] **T.-Y. Lin, P. Goyal, R. Girshick, K. He und P. Dollar**. *Focal loss for dense object detection*. In: *IEEE International Conference on Computer Vision*. 2017, S. 2980–2988.

[79] **J.-Y. Liu und Y.-H. Yang**. *Event localization in music auto-tagging*. In: *24th ACM International Conference on Multimedia*. 2016, S. 1048–1057.

[80] **J.-Y. Liu und Y.-H. Yang**. *Dilated convolution with dilated GRU for music source separation*. In: *International Joint Conference on Artificial Intelligence*. 2019, S. 4718–4724.

[81] **P. C. Loizou**. *Speech quality assessment*. In: *Multimedia Analysis, Processing and Communications*. Hrsg. von **W. Lin, D. Tao, J. Kacprzyk, Z. Li, E. Izquierdo und H. Wang**. Springer Berlin, Heidelberg, 2011, S. 623–654.

[82] **H. Löllmann, E. Yilmaz, M. Jeub und P. Vary**. *An improved algorithm for blind reverberation time estimation*. In: *International Workshop on Acoustic Echo and Noise Control*. 2010, S. 1–4.

[83] **Y. Luo und J. Yu**. *Music source separation with band-split RNN*. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), S. 1893–1901.

[84] **E. Manilow, P. Seetharaman und B. Pardo**. *Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2020, S. 771–775.

[85] **E. Manilow, G. Wichern und J. Le Roux**. *Hierarchical musical instrument separation*. In: *21st International Society for Music Information Retrieval Conference*. 2020, S. 376–383.

[86] **E. Manilow, G. Wichern, P. Seetharaman und J. Le Roux**. *Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity*. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2019, S. 45–49.

[87] **M. Mauch und S. Ewert**. *The audio degradation toolbox and its application to robustness evaluation*. In: *14th International Society for Music Information Retrieval Conference*. 2013, S. 83–88.

[88] **W. S. McCulloch und W. Pitts**. *A logical calculus of the ideas immanent in nervous activity*. In: *The Bulletin of Mathematical Biophysics* 5.4 (1943), S. 115–133.

[89] **B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg und O. Nieto**. *librosa: Audio and music signal analysis in python*. In: *14th python in science conference*. 2015, S. 18–25.

[90] **G. Meseguer-Brocal und G. Peeters**. *Content based singing voice source separation via strong conditioning using aligned phonemes*. In: *21st International Society for Music Information Retrieval Conference*. 2020, S. 819–827.

[91] **G. Meseguer-Brocal und G. Peeters**. *Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations*. In: *20th International Society for Music Information Retrieval Conference*. 2019, S. 159–165.

[92] **MIDI Association**. *General MIDI 1 Sound Set*. URL: https://www.midi.org/specifications-old/item/gm-level-1-sound-set (besucht am 01. 02. 2023).

[93] **M. Miron, J. Janer und E. Gómez**. *Generating data to train convolutional neural networks for low latency classical music source separation*. In: *14th Sound and Music Computing Conference*. 2017, S. 227–233.

[94] **M. Miron, J. Janer und E. Gómez**. *Monaural score-informed source separation for classical music using convolutional neural networks*. In: *18th International Society for Music Information Retrieval Conference*. 2017, S. 55–62.

[95] **T. M. Mitchell**. *Machine learning*. 9. McGraw-Hill New York, 1997.

[96] **M. Möser**. *Technische Akustik*. Springer Berlin, Heidelberg, 2012.

[97] **R. Moussaoui, J. Rouat und R. Lefebvre**. *Wavelet based independent component analysis for multi-channel source separation*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Bd. 5. 2006, S. V645–V648.

[98] **M. Müller**. *Fundamentals of music processing: Using Python and Jupyter notebooks*. Springer Cham, 2021.

[99] **National Institute of Standards and Technology**. *NIST speech signal to noise ratio measurements*. URL: https://www.nist.gov/itl/iad/mig/nist-speech-signal-noise-ratio-measurements (besucht am 19. 01. 2023).

[100] **T. Necciari, P. Balazs, N. Holighaus und P. L. Søndergaard**. *The ERBlet transform: An auditory-based time-frequency representation with perfect reconstruction*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, S. 498–502.

[101] **A. Ng**. *Machine learning yearning: Technical strategy for AI engineers in the era of deep learning*. E-Print: deeplearning.ai, 2018.

[102] **L. Pandey, A. Kumar und V. Namboodiri**. *Monaural audio source separation using variational autoencoders*. In: *Interspeech*. 2018, S. 3489–3493.

[103] **P. Papadopoulos, R. Travadi und S. S. Narayanan**. *Global SNR estimation of speech signals for unknown noise conditions using noise adapted non-linear regression*. In: *Interspeech*. 2017, S. 3842–3846.

[104] **J. Pons, O. Slizovskaia, R. Gong, E. Gómez und X. Serra**. *Timbre analysis of music audio signals with convolutional neural networks*. In: *25th European Signal Processing Conference*. 2017, S. 2744–2748.

[105] **F. Puente León und H. Jäkel**. *Signale und Systeme*. De Gruyter Oldenbourg, 2019.

[106] **C. Raffel und D. P. W. Ellis**. *Intuitive analysis, creation and manipulation of MIDI data with pretty_midi*. In: *15th International Conference on Music Information Retrieval Late Breaking and Demo Papers*. 2014.

[107] **C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang und D. P. W. Ellis**. *mir_eval: A transparent implementation of common MIR metrics*. In: *15th International Society for Music Information Retrieval Conference*. 2014.

[108] **Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis und R. Bittner**. *MUSDB18 - a corpus for music separation*. Version 1.0.0. Zenodo, 2017.

[109] **Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, D. FitzGerald und B. Pardo**. *An overview of lead and accompaniment separation in music*. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.8 (2018), S. 1307–1335.

[110] **A. W. Rix, J. G. Beerends, M. P. Hollier und A. P. Hekstra**. *Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Bd. 2. 2001, S. 749–752.

[111] **O. Ronneberger, P. Fischer und T. Brox**. *U-Net: Convolutional networks for biomedical image segmentation*. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2015, S. 234–241.

[112] **S. Rouard, F. Massa und A. Défossez**. *Hybrid transformers for music source separation*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2023, S. 1–5.

[113] **W. C. Sabine**. *Collected papers on acoustics*. Cambridge Harvard University Press, 1922.

[114] **J. Salamon, C. Jacoby und J. P. Bello**. *A dataset and taxonomy for urban sound research*. In: *22nd ACM international conference on Multimedia*. 2014, S. 1041–1044.

[115] **L. Savioja und U. P. Svensson**. *Overview of geometrical room acoustic modeling techniques*. In: *The Journal of the Acoustical Society of America* 138.2 (2015), S. 708–730.

[116] **N. Scaringella, G. Zoia und D. Mlynek**. *Automatic genre classification of music content: a survey*. In: *IEEE Signal Processing Magazine* 23.2 (2006), S. 133–141.

[117] **C. Schörkhuber und A. Klapuri**. *Constant-Q transform toolbox for music processing*. In: *7th Sound and Music Computing Conference*. 2010, S. 3–64.

[118] **C. Schörkhuber, A. Klapuri, N. Holighaus und M. Dörfler**. *A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution*. In: *53rd International Conference of the Audio Engineering Society on Semantic Audio*. 2014.

[119] **M. R. Schroeder**. *New method of measuring reverberation time*. In: *The Journal of the Acoustical Society of America* 37.6 (1965), S. 1187–1188.

[120] **M. Schuster und K. K. Paliwal**. *Bidirectional recurrent neural networks*. In: *IEEE Transactions on Signal Processing* 45.11 (1997), S. 2673–2681.

[121] **J. Sebastian und H. A. Murthy**. *Group delay based music source separation using deep recurrent neural networks*. In: *International Conference on Signal Processing and Communications*. 2016, S. 1–5.

[122] **R. Serizel, N. Turpault, A. Shah und J. Salamon**. *Sound event detection in synthetic domestic environments*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2020, S. 86–90.

[123] **T. Sgouros, A. Bousis und N. Mitianoudis**. *An efficient short-time discrete cosine transform and attentive MultiResUNet framework for music source separation*. In: *IEEE Access* 10 (2022), S. 119448–119459.

[124] **K. Simonyan und A. Zisserman**. *Very deep convolutional networks for large-scale image recognition*. In: *3rd International Conference on Learning Representations*. 2015.

[125] **O. Slizovskaia, G. Haro und E. Gómez**. *Conditioned source separation for musical instrument performances*. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), S. 2083–2095.

[126] **O. Slizovskaia, L. Kim, G. Haro und E. Gómez**. *End-to-end sound source separation conditioned on instrument labels*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2019, S. 306–310.

[127] **P. Smaragdis und J. C. Brown**. *Non-negative matrix factorization for polyphonic music transcription*. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2003, S. 177–180.

[128] **N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever und R. Salakhutdinov**. *Dropout: A simple way to prevent neural networks from overfitting*. In: *The Journal of Machine Learning Research* 15.1 (2014), S. 1929–1958.

[129] **G.-B. Stan, J.-J. Embrechts und D. Archambeau**. *Comparison of different impulse response measurement techniques*. In: *Journal of the Audio Engineering Society* 50.4 (2002), S. 249–262.

[130] **S. Stevens, J. Volkmann und E. Newman**. *A scale for the measurement of the psychological magnitude pitch*. In: *The Journal of the Acoustical Society of America* 8.3 (1937), S. 185–190.

[131] **D. Stoller, S. Ewert und S. Dixon**. *Wave-U-Net: A multi-scale neural network for end-to-end audio source separation*. In: *19th International Society for Music Information Retrieval Conference*. 2018, S. 334–340.

[132] **F.-R. Stöter, S. Uhlich, A. Liutkus und Y. Mitsufuji**. *Open-Unmix – A reference implementation for music source separation*. In: *Journal of Open Source Software* 4.41, 1667 (2019).

[133] **V. Subramanian, E. Benetos und M. Sandler**. *Robustness of adversarial attacks in sound event classification*. In: *4th Workshop on Detection and Classification of Acoustic Scenes and Events*. 2019, S. 239–243.

[134] **I. Sutskever, O. Vinyals und Q. V. Le**. *Sequence to sequence learning with neural networks*. In: *Advances in Neural Information Processing Systems*. Bd. 27. 2014.

[135] **I. Szöke, M. Skácel, L. Mošner, J. Paliesek und J. H. Černockỳ**. *Building and evaluation of a real room impulse response dataset*. In: *IEEE Journal of Selected Topics in Signal Processing* 13.4 (2019), S. 863–876.

[136] **N. Takahashi und Y. Mitsufuji**. *D3net: Densely connected multi-dilated densenet for music source separation*. In: *arXiv preprint arXiv:2010.01733* (2020).

[137] **J. Thickstun, Z. Harchaoui und S. M. Kakade**. *Learning features of music from scratch*. In: *International Conference on Learning Representations*. 2017.

[138] **A. Uemura, K. Ishikura und J. Katto**. *Effects of audio compression on chord recognition*. In: *International Conference on Multimedia Modeling*. 2014, S. 345–352.

[139] **S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi und Y. Mitsufuji**. *Improving music source separation based on deep neural networks through data augmentation and network blending*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2017, S. 261–265.

[140] **L. Vande Veire und T. De Bie**. *From raw audio to a seamless mix: creating an automated DJ system for drum and bass*. In: *EURASIP Journal on Audio, Speech, and Music Processing* (2018), S. 1–21.

[141] **A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser und I. Polosukhin**. *Attention is all you need*. In: *Advances in Neural Information Processing Systems*. Bd. 30. 2017.

[142] **G. Velarde, T. Weyde, C. C. Chacón, D. Meredith und M. Grachten**. *Composer recognition based on 2D-filtered piano-rolls*. In: *17th International Society for Music Information Retrieval Conference*. 2016, S. 115–121.

[143] **E. Vincent, R. Gribonval und C. Févotte**. *Performance measurement in blind audio source separation*. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.4 (2006), S. 1462–1469.

[144] **E. Vincent, N. Bertin, R. Gribonval und F. Bimbot**. *From blind to guided audio source separation: How models and side information can improve the separation of sound*. In: *IEEE Signal Processing Magazine* 31.3 (2014), S. 107–115.

[145] **P. Vincent, H. Larochelle, Y. Bengio und P.-A. Manzagol**. *Extracting and composing robust features with denoising autoencoders*. In: *25th International Conference on Machine Learning*. 2008, S. 1096–1103.

[146] **A. C. Vinci**. *Die Notenschrift: Grundlagen der traditionellen Musiknotation*. Bärenreiter-Verlag, 1988.

[147] **T. Virtanen**. *Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria*. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.3 (2007), S. 1066–1074.

[148] **T. Virtanen und T. Heittola**. *Interpolating hidden Markov model and its application to automatic instrument recognition*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009, S. 49–52.

[149] **S. Weinzierl**, Hrsg. *Handbuch der Audiotechnik*. Springer Berlin, Heidelberg, 2008.

[150] **Z. Wu, X. Xiao, E. S. Chng und H. Li**. *Synthetic speech detection using temporal modulation feature*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, S. 7234–7238.

[151] **H. Yang, S. Firodiya, N. J. Bryan und M. Kim**. *Don't separate, learn to remix: end-to-end neural remixing with joint optimization*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2022, S. 116–120.

[152] **A. Zadeh, T. Ma, S. Poria und L.-P. Morency**. *Wildmix dataset and spectro-temporal transformer model for monoaural audio source separation*. In: *arXiv preprint arXiv:1911.09783* (2019).

[153] **D. Zhu und K. K. Paliwal**. *Product of power spectrum and group delay function for speech recognition*. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Bd. 1. 2004, S. I125–I128.

Eigene Veröffentlichungen

[A1] **A. Kacaras, M. Bächle, M. Schwabe, F. Zanger, F. Puente León und V. Schulze**. *Acoustic emission-based characterization of focal position during ultra-short pulse laser ablation*. In: *Procedia CIRP* 81 (2019), S. 270–275.

[A2] **M. Schwabe, O. Elaiashy und F. Puente León**. *Incorporation of phase information for improved time-dependent instrument recognition*. In: *tm - Technisches Messen* 87.s1 (2020), s62–s67.

[A3] **M. Schwabe und M. Heizmann**. *Influence of input data representations for time-dependent instrument recognition*. In: *tm - Technisches Messen* 88.5 (2021), S. 274–281.

[A4] **M. Schwabe und M. Heizmann**. *Improved separation of polyphonic chamber music signals by integrating instrument activity labels*. In: *IEEE Access* 11 (2023), S. 42999–43007.

[A5] **M. Schwabe, S. Murgul und M. Heizmann**. *Dual task monophonic singing transcription*. In: *Journal of the Audio Engineering Society* 70.12 (2022), S. 1038–1047.

[A6] **M. Schwabe, M. Weber und F. Puente León**. *Notenseparation in polyphonen Musiksignalen durch einen Matching-Pursuit-Algorithmus*. In: *tm - Technisches Messen* 85.s1 (2018), s103–s109.

[A7] **M. Schwabe, T. Hoffmann, S. Murgul und M. Heizmann**. *Estimation of music recording quality to predict automatic music transcription performance*. In: *3rd International Conference on Smart Multimedia*. 2022, S. 322–337.

[A8] **F. Zanger, A. Kacaras, M. Bächle, M. Schwabe, F. Puente León und V. Schulze**. *FEM simulation and acoustic emission based characterization of chip segmentation frequency in machining of Ti-6Al-4V*. In: *Procedia CIRP* 72 (2018), S. 1421–1426.

Betreute studentische Arbeiten

[B1] **B. Bau**. *Weiterentwicklung eines NMF-Algorithmus und Aufbau einer Testumgebung für die Transkription polyphoner Musiksignale*. Masterarbeit. KIT, 2019.

[B2] **R. S. Beeh**. *Separation of music signals of different musical instruments with Transformer networks*. Bachelorarbeit. KIT, 2022.

[B3] **K. Chang**. *Transcription of polyphonic melodies with neural networks*. Bachelorarbeit. KIT, 2019.

[B4] **C. Ding**. *Variational Autoencoder for direct time signal separation of different music instruments*. Masterarbeit. KIT, 2020.

[B5] **O. Elaiashy**. *Integration der Phaseninformation zur verbesserten Instrumentenerkennung in Musikstücken*. Bachelorarbeit. KIT, 2019.

[B6] **H. Firzlaff**. *Estimation and evaluation of recording quality of polyphonic music signals with various instruments*. Bachelorarbeit. KIT, 2021.

[B7] **Y. Griri**. *Development of Variational Autoencoder enhancements for the separation of music signals of different instruments*. Bachelorarbeit. KIT, 2020.

[B8] **T. Hoffmann**. *Automatische Qualitätsschätzung von Musiksignalen*. Masterarbeit. KIT, 2020.

[B9] **J. Hraman**. *Separation of band music with neural networks*. Masterarbeit. KIT, 2018.

[B10] **R. Ji**. *Direct component separation for music recordings from music bands using neural networks*. Masterarbeit. KIT, 2020.

[B11] **S. Li**. *Comparison of autoencoder structures for component separation of band music recordings*. Masterarbeit. KIT, 2020.

[B12] **E. J. Lozano Corredor**. *Augmentierung mithilfe realitätsnaher synthetischer Musiksignale für die Separation von realen Instrumentenaufnahmen*. Bachelorarbeit. KIT, 2022.

[B13] **A. Márquez González**. *Integration of information about active instruments in the separation of polyphonic music signals*. Masterarbeit. KIT, 2021.

[B14] **S. Murgul**. *Automated transcription of art song audio signals*. Masterarbeit. KIT, 2020.

[B15] **R. R. P. Nugroho**. *Recurrent neural networks for component separation in band music*. Bachelorarbeit. KIT, 2019.

[B16] **C. Park**. *Separation der harmonischen und perkussiven Anteile polyphoner Musiksignale*. Bachelorarbeit. KIT, 2018.

[B17] **K. Y. Qiu**. *Transcription of polyphonic melodies using non-negative matrix factorization (NMF)*. Bachelorarbeit. KIT, 2018.

[B18] **M. Reiser**. *Transkription polyphoner Musiksignale mit Methoden des maschinellen Lernens*. Masterarbeit. KIT, 2019.

[B19] **A. Schnerring**. *Schätzung instrumentenspezifischer Notenspektren zur Musikseparation mithilfe neuronaler Netze*. Bachelorarbeit. KIT, 2019.

[B20] **K. Triki**. *Time-dependent music instrument recognition without preceding signal transformation*. Bachelorarbeit. KIT, 2021.

[B21] **M. Wachter**. *Zeitabhängige Instrumentenerkennung in Musikstücken mittels Betrags- und Phasenspektren*. Bachelorarbeit. KIT, 2020.

[B22] **M. P. Weber**. *Separation von polyphonen Melodien durch einen Matching-Pursuit-Algorithmus*. Masterarbeit. KIT, 2018.

[B23] **J. Zhao**. *Generation of realistic music data for an improved generalization in the separation of polyphonic music signals*. Masterarbeit. KIT, 2021.

Forschungsberichte
aus der Industriellen Informationstechnik
(ISSN 2190-6629)

Band 1 Pérez Grassi, Ana
Variable illumination and invariant features for detecting and classifying varnish defects.
ISBN 978-3-86644-537-6

Band 2 Christ, Konrad
Kalibrierung von Magnet-Injektoren für Benzin-Direkteinspritzsysteme mittels Körperschall.
ISBN 978-3-86644-718-9

Band 3 Sandmair, Andreas
Konzepte zur Trennung von Sprachsignalen in unterbestimmten Szenarien.
ISBN 978-3-86644-744-8

Band 4 Bauer, Michael
Vergleich von Mehrträger-Übertragungsverfahren und Entwurfskriterien für neuartige Powerline-Kommunikationssysteme zur Realisierung von Smart Grids.
ISBN 978-3-86644-779-0

Band 5 Kruse, Marco
Mehrobjekt-Zustandsschätzung mit verteilten Sensorträgern am Beispiel der Umfeldwahrnehmung im Straßenverkehr.
ISBN 978-3-86644-982-4

Band 6 Dudeck, Sven
Kamerabasierte In-situ-Überwachung gepulster Laserschweißprozesse.
ISBN 978-3-7315-0019-3

Band 7 Liu, Wenqing
Emulation of Narrowband Powerline Data Transmission Channels and Evaluation of PLC Systems.
ISBN 978-3-7315-0071-1

Die Bände sind unter www.ksp.kit.edu als PDF frei verfügbar oder als Druckausgabe bestellbar.

Band 8 Otto, Carola
Fusion of Data from Heterogeneous Sensors with Distributed Fields of View and Situation Evaluation for Advanced Driver Assistance Systems.
ISBN 978-3-7315-0073-5

Band 9 Wang, Limeng
Image Analysis and Evaluation of Cylinder Bore Surfaces in Micrographs.
ISBN 978-3-7315-0239-5

Band 10 Michelsburg, Matthias
Materialklassifikation in optischen Inspektionssystemen mithilfe hyperspektraler Daten.
ISBN 978-3-7315-0273-9

Band 11 Pallauf, Johannes
Objektsensitive Verfolgung und Klassifikation von Fußgängern mit verteilten Multi-Sensor-Trägern.
ISBN 978-3-7315-0529-7

Band 12 Sigle, Martin
Robuste Schmalband-Powerline-Kommunikation für Niederspannungsverteilernetze.
ISBN 978-3-7315-0539-6

Band 13 Opalko, Oliver
Powerline-Kommunikation für Batteriemanagement-Systeme in Elektro- und Hybridfahrzeugen.
ISBN 978-3-7315-0647-8

Band 14 Han, Bin
Characterization and Emulation of Low-Voltage Power Line Channels for Narrowband and Broadband Communication.
ISBN 978-3-7315-0654-6

Band 15 Alonso, Damián Ezequiel
Wireless Data Transmission for the Battery Management System of Electric and Hybrid Vehicles.
ISBN 978-3-7315-0670-6

Band 16 Hernández Mesa, Pilar
Design and analysis of a content-based image retrieval system.
ISBN 978-3-7315-0692-8

Band 17 Suchaneck, André
Energiemanagement-Strategien für batterieelektrische Fahrzeuge.
ISBN 978-3-7315-0773-4

Band 18 Bauer, Sebastian
Hyperspectral Image Unmixing Incorporating Adjacency Information.
ISBN 978-3-7315-0788-8

Band 19 Vater, Sebastian
Monokulare Blickrichtungsschätzung zur berührungslosen Mensch-Maschine-Interaktion.
ISBN 978-3-7315-0789-5

Band 20 Back, Kristine
Erkennung menschlicher Aktivitäten durch Erfassung und Analyse von Bewegungstrajektorien.
ISBN 978-3-7315-0909-7

Band 21 Nürnberg, Thomas
Entwurf von Computational-Imaging-Systemen am Beispiel der monokularen Tiefenschätzung.
ISBN 978-3-7315-0941-7

Band 22 Kaiser, Cornelius
Adaptive Modulationsverfahren für die schmalbandige Powerline-Kommunikation in Niederspannungsnetzen.
ISBN 978-3-7315-1010-9

Band 23 Struckmeier, Frederick
Prozesssicherheit von Laserschneidmaschinen – Auflagemessung und Schachtelung
ISBN 978-3-7315-1127-4

Band 24 Tatzel, Leonie Felica
Verbesserungen beim Laserschneiden mit Methoden des maschinellen Lernens
ISBN 978-3-7315-1128-1

Die Bände sind unter www.ksp.kit.edu als PDF frei verfügbar oder als Druckausgabe bestellbar.

Band 25 Wetzel, Johannes
Probabilistic Models and Inference for Multi-View People Detection in Overlapping Depth Images
ISBN 978-3-7315-1177-9

Band 26 Mitschke, Norbert
Konvolutionäre neuronale Netze in der industriellen Bildverarbeitung und Robotik
ISBN 978-3-7315-1197-7

Band 27 Schambach, Maximilian
Reconstruction from Spatio-Spectrally Coded Multispectral Light Fields
ISBN 978-3-7315-1210-3

Band 28 Bächle, Matthias
Model-based Filtering of Interfering Signals in Ultrasonic Time Delay Estimations
ISBN 978-3-7315-1252-3

Band 29 Demirdelen, Ismet
Nutzung der keramischen Glühkerze als Sensorelement
ISBN 978-3-7315-1265-3

Band 30 Anastasiadis, Johannes
Überwachte Methoden für die spektrale Entmischung mit künstlichen neuronalen Netzen
ISBN 978-3-7315-1305-6

Band 31 Uhlig, David
Light Field Imaging for Deflectometry
ISBN 978-3-7315-1306-3

Band 32 Hartung, Julia
Machine Learning for Camera-Based Monitoring of Laser Welding Processes
ISBN 978-3-7315-1333-9

Band 33 Li, Lanxiao
Computational, Label, and Data Efficiency in Deep Learning for Sparse 3D Data
ISBN 978-3-7315-1346-9

Band 34 Schwabe, Markus
Analyse und Separation polyphoner Musiksignale
ISBN 978-3-7315-1365-0

ISSN 2190-6629
ISBN 978-3-7315-1365-0

Gedruckt auf FSC-zertifiziertem Papier