Technische Universität Dortmund
Fakultät Statistik

Musikklassifikation mittels auditorischer Modelle zur Optimierung von Hörgeräten

---

Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften

von

Dipl.-Inf.
KLAUS FRIEDRICHS

Vorgelegt: Dortmund, 30. Mai 2016
Tag der mündlichen Prüfung: 22. Juni. 2016
Betreuer: Prof. Dr. Claus Weihs
Zweitgutachter: Dr. Uwe Ligges
Kommissionsvorsitz: Prof. Dr. Jörg Rahnenführer

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Dissertation ist bisher keiner anderen Fakultät vorgelegt worden. Ich erkläre, dass ich bisher kein Promotionsverfahren erfolglos beendet habe und dass keine Aberkennung eines bereits erworbenen Doktorgrades vorliegt.

Klaus Friedrichs

Inhaltsverzeichnis

1 Einleitung

Laut einer Studie sind in Deutschland 19% der Bevölkerung über 14 Jahren hörbeeinträchtigt, wobei dieser Anteil auf Grund des demografischen Wandels noch weiter zunimmt (Sohn, 2000). Für diesen Personenkreis ist neben Beeinträchtigungen im täglichen Leben, wie beispielsweise Schwierigkeiten bei der Lokalisation von Geräuschen oder bei der Sprachverständlichkeit, auch der Hörgenuss von Musik eingeschränkt. Diese wird häufig als verrauscht wahrgenommen, wobei vor allem die Tonhöhen- und die Klangfarbenwahrnehmung beeinträchtigt sind (Feldmann und Kumpf, 1988; McDermott, 2004; Gfeller u. a., 2006). Für viele Hörschädigungen können diese Effekte durch moderne digitale Hörgeräte gemindert werden, die hörschädigungsspezifisch die ankommenden akustischen Signale modifizieren. Während die Hörgeräteentwicklung lange Zeit fast ausschließlich auf die Verbesserung des Sprachverständnisses fokussiert war, ist erst vor wenigen Jahren das Interesse an der Verbesserung der Musikwahrnehmung gestiegen. Dazu haben auch Forschungsprojekte, wie das Teilprojekt „Statistische Modellierung zeitlich und spektral hoch aufgelöster Audiodaten in Hörgeräten“ des SFB 823, in dessen Rahmen diese Arbeit entstanden ist, beigetragen.

Sowohl für die Sprach- als auch für die Musikverbesserung besteht bei der Entwicklung und Optimierung von Hörgerätealgorithmen das Hindernis, dass diese üblicherweise nur mit Hilfe von aufwändigen Hörversuchen validiert werden können. Dabei ist zudem die Probandenzahl meist sehr klein (ca. 20 Probanden ist ein typischer Wert), obwohl sich Hörschädigungen sehr unterschiedlich auswirken und daher oft die optimalen Einstellungen eines Hörgerätealgorithmus nicht für alle Probanden gleich sind. Weiterhin können bei der Optimierung mit Hörversuchen nur wenige Varianten eines Algorithmus untersucht werden und diese müssen auch schon im Vorfeld definiert sein, was eine sequentielle Optimierung, die auf die Bewertung der bereits untersuchten Varianten eingeht, unmöglich macht. Wünschenswert ist daher ein Verfahren, das für einen Hörgerätealgorithmus automatisch (ohne Hörversuch) evaluiert, wie gut damit bei einer gegebenen Hörschädigung Sprache oder Musik wahrgenommen werden kann. In diese Arbeit wird ein solches Verfahren für

Musik entwickelt, das für einen simulierten Hörer die Fähigkeit zur Differenzierung von Musikeigenschaften misst.

Als Basis für die Substitution des menschlichen Hörers bietet sich ein Simulationsmodell des menschlichen Hörvorgangs an. In den letzten Jahrzehnten sind verschiedene sogenannte auditorische Modelle entwickelt worden. Diese simulieren computerbasiert die verschiedenen Stufen des Hörvorgangs, beginnend mit der Schallaufnahme im Ohr – auch als auditorische Peripherie bezeichnet – bis hin zur Informationsextraktion durch die Hörrinde im Großhirn. Dabei ist die Detailschärfe der verschiedenen Modelle jedoch sehr unterschiedlich ausgeprägt, und es existiert auch noch keine Theorie, die alles relevante physiologische und psychologische Wissen integriert (Meddis u. a., 2010b). Relativ verlässlich ist bislang lediglich die auditorische Peripherie erforscht, die ausgehend von den eintreffenden Schallwellen im Außenohr, die Transformationen im Ohr bis hin zu den ausgehenden Nervenimpulsen im Hörnerv beschreiben. Für diesen Verarbeitungsschritt gibt es detailgetreue Modelle, die umfangreich mittels psychoakustischer Experimente und Beobachtungen an Tieren validiert sind (z.B. Meddis und O'Mard (2005) und Lopez-Najera, Lopez-Poveda und Meddis (2007)). Eine Funktionsstörung innerhalb der auditorischen Peripherie, die sogenannte sensorineurale Schwerhörigkeit, ist zudem auch die meist verbreitetste Form der Hörschädigungenen. In mehreren Studien wurde bereits gezeigt, dass es möglich ist, sensorineurale Schwerhörigkeiten realitätsnah in Modelle der auditorischen Peripherie – auch kurz als Ohrmodell bezeichnet – einzubauen (Heinz u. a., 2001; Zilany und Bruce, 2006; Jepsen und Dau, 2011; Panda u. a., 2014).

Für ein akustisches Signal, bestehend aus M Abtastwerten, ist die Ausgabe des Ohrmodells ein multivariates Signal der Größe $K \times M$, wobei K die Anzahl der simulierten Hörnervenfasern bezeichnet. Jede Hörnervenfaser reagiert besonders stark auf eine bestimmte Frequenzregion, deren Zentrum als *Best Frequenz* (oder auch charakteristische Frequenz) bezeichnet wird (Roederer und Mayer, 1999). In Abbildung 1.1 ist ein Beispiel für die Ausgabe eines Ohrmodells, sowohl mit als auch ohne simulierte Hörschädigung abgebildet. Die Farbe kennzeichnet, wie viele Spikes (Nervenimpulse) die Hörnervenfasern, die entsprechend ihrer *Best Frequenzen* auf der y-Achse angeordnet sind, zu einem bestimmten Zeitpunkt feuern. In dem Beispiel mit Hörschädigung erkennt man eine deutlich niedrigere Aktivität, die sich in diesem Fall besonders stark auf die hohen Frequenzen auswirkt.

Es ist somit möglich die Ausgabe eines Modells mit Hörschädigung mit der Ausgabe

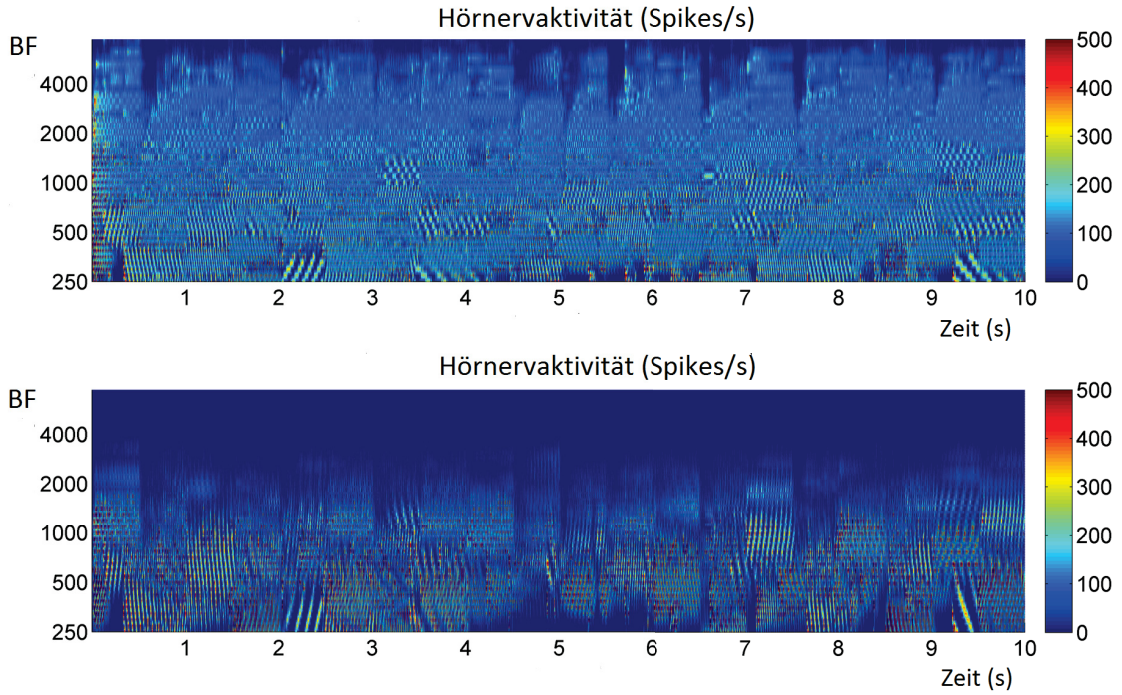

Abbildung 1.1: Beispielhafte Ohrmodellausgabe für das Modell von Meddis (2006) ohne Hörschädigung (oben) und mit einer Hörschädigung (unten). Die Farbe signalisiert die Aktivität der Hörnervenfasern entsprechend ihrer *Best Frequenzen* (BF) im Zeitverlauf.

des Modell ohne Hörschädigung zu vergleichen.[1] Eine einfache Möglichkeit dafür ist die Verwendung eines quantitativen Distanzmaßes, wie beispielsweise die mittleren Abstände über alle Punkte (z.B. *Root-Mean-Square Error*, RMSE) oder das in Hines und Harte (2010) für die Bewertung der Sprachverständlichkeit vorgeschlagene Maß MSSIM (*Mean Structured Similarity Index*). Dieses Maß berechnet fensterweise einen Wert $SSIM(x, y)$, der sich aus den Unterschieden der Mittelwerte und Varianzen der beiden Ausgaben x und y, sowie deren Korrelationskoeffizienten zusammensetzt, und bildet anschließend den Mittelwert über alle Fenster.

Das Problem bei der Verwendung all dieser Distanzmaße für die Optimierung eines Hörgerätealgorithmus ist allerdings, dass dadurch die Hörnervenaktivität des Normalhörenden als Ziel definiert wird, was aus verschiedenen Gründen oft nicht richtig ist. Beispielsweise kann eine Rauschreduktion oder eine andere Art der Komplexitätsreduktion sinnvoll

[1] In dieser Arbeit werden Modelle, die eine Hörschädigung simulieren, zur besseren Lesbarkeit als „Modelle mit Hörschädigung“ bezeichnet und das Originalmodell ohne simulierte Hörschädigung als „Modell ohne Hörschädigung“.

sein, obwohl dies eine höhere Distanz zur Folge hat, z.B. bei einer Hörschädigung mit beeinträchtigter Frequenzauflösung. Generell kann ein solches Distanzmaß nicht bewerten, welche Information wie wichtig ist. Praktisch sind alle Informationen, die durch die Feuerraten der Hörnervenfasern kodiert sind, redundant in mehreren Fasern enthalten. Ein weiterer Punkt, der gegen ein einfaches Distanzmaß spricht, ist das Szenario, in dem der Ausfall von Haarzellen durch die Verstärkung der Aktivität von anderen Haarzellen kompensiert werden kann. Ein Beispiel dafür ist ein Hörgerätealgorithmus, der die Frequenzen eines Signals so verändert, dass diese in einen besser wahrnehmbaren Bereich verschoben werden. Zusammenfassend lässt sich sagen, ein quantitatives Distanzmaß wie RMSE oder MSSIM kann helfen den Grad einer Schwerhörigkeit bezüglich eines bestimmten Signals abzuschätzen, es ist aber weniger dafür geeignet einen Hörgerätealgorithmus zu bewerten.

Was man stattdessen benötigt, ist ein qualitatives Distanzmaß, das den aus der Hörschädigung resultierenden Informationsverlust qualitativ bewertet. Unter der Annahme, dass trotz einiger Vereinfachungen das Ohrmodell die Realität hinreichend genau abbildet, muss die musikalische (oder sonstige) Information eines Signals vollständig in der Ohrmodellausgabe kodiert sein. Bei einem Modell mit Hörschädigung ist diese Kodierung beschädigt, wodurch ein Teil der Information nicht mehr eindeutig identifizierbar ist. In der Realität führt dies dazu, dass Hörgeschädigte Musik häufig als verrauscht wahrnehmen. Wie genau die akustische Information im Gehirn extrahiert wird, ist allerdings noch lange nicht vollständig geklärt. Als Ersatz bieten sich jedoch statistische Klassifikationsverfahren an, die ähnlich wie das Gehirn nach typisch wiederkehrenden Mustern suchen.

Das Ziel dieser Arbeit ist die Entwicklung einer auf der Ohrmodellausgabe aufbauenden Methode, die mit Hilfe von statistischen Klassifikationsverfahren die Erkennungsraten der relevanten Musikeigenschaften schätzt, um somit qualitativ Hörschädigungen und Hörgeräte bewerten zu können. Ähnliche Bewertungsverfahren werden derzeit auch für Sprache entwickelt (Jürgens u. a., 2014; Karbasi und Kolossa, 2015), wobei dort jedoch die Zielfunktion trivialer ist, denn Sprachverständlichkeit lässt sich relativ intuitiv durch die Spracherkennungsrate definieren. Diese Verfahren verwenden demzufolge einen automatischen Spracherkenner, der auf der Ausgabe des Ohrmodells aufbaut. Die Güte eines Hörgeräts bezüglich einer spezifischen, simulierten Hörschädigung kann dann einfach durch die Fehlerrate des Verfahrens bestimmt werden. Prinzipiell funktioniert auch das in dieser Arbeit entwickelte Bewertungsverfahren für Musik analog, indem statt der Sprachverständlichkeit die Musikverständlichkeit geschätzt wird. Jedoch muss hierfür erst

einmal definiert werden, was Musikverständlichkeit überhaupt ist. Laut Fitz und McKinney (2015) gibt es keine direkte Methode diese zu ermitteln und stattdessen müssen die Einzelbestandteile Tonhöhe, Harmonie, Klangfarbe und Rhythmus unabhängig ermittelt werden. Gebräuchlicher ist allerdings eine Definition der Musikwahrnehmung durch die vier Einzelbestandteile Tonhöhe, Lautheit (wahrgenommene Lautstärke), Klangfarbe und Tondauer (Rhythmus) (Roederer und Mayer, 1999; Eronen, 2001), die daher auch hier verwendet wird. Ein sinnvoller Ansatz zur Messung der Musikverständlichkeit ist somit, die Erkennungsraten dieser vier Attribute unabhängig zu ermitteln und anschließend entweder zu einem Gesamtmaß zu aggregieren oder das Ergebnis mehrkriteriell zu betrachten. Allerdings sind die Attribute subjektiv und müssten zunächst durch Hörversuche mit Normalhörenden für einen Musikdatensatz erfasst werden, was im Rahmen dieser Arbeit jedoch nicht möglich ist. Stattdessen werden MIDI-Stücke verwendet und die darin enthaltene objektive Information der Musik (Tonanfänge, Tondauer, Musikinstrumente, sowie die physikalischen Tonhöhen und Lautstärken) mit psychoakustischen Erkenntnissen kombiniert. Die genaue Umsetzung für die einzelnen Attribute wird im Folgenden beschrieben.

Die wahrgenommene Tonhöhe ist eng verknüpft mit der Grundfrequenz, wobei jedoch relative Frequenzunterschiede unter 3% (ein halber Halbton) auch von Normalhörenden oft nicht erkannt werden können (Roederer und Mayer, 1999). In dieser Arbeit wird daher die Tonhöhenwahrnehmung durch eine Grundfrequenzschätzung simuliert, wobei eine Fehlertoleranz von 3% zugelassen wird. Die Lautheit eines Tons hängt natürlich von der physikalischen Lautstärke ab, aber auch von der Tonhöhe, der Tondauer und möglichen Überlagerungen mit anderen Tönen. Eine direkte mathematische Ableitung von der Lautstärke auf die Lautheit existiert bislang nur für reine Töne (siehe Kapitel 2.2), aber nicht für komplexe Töne und erst recht nicht für polyphone Musik. Für das in dieser Arbeit entwickelte Bewertungsverfahren wird daher die Lautheit nicht berücksichtigt. Die Schätzung der Tondauer wird durch eine Schätzung der Toneinsatzzeiten ersetzt, da es für diese Aufgabe bereits umfangreiche Forschungsarbeiten gibt. Für einstimmige Musikstücke ohne Pausen folgt aus der Schätzung der Einsatzzeiten auch direkt die Tondauer. In Zukunft könnte das Verfahren noch durch eine Pausenerkennung erweitert werden. Die Klangfarbe (auch als Timbre bezeichnet) ist definiert, als die Eigenschaften eines Tons, die es einen Hörer ermöglicht zwei Töne der gleichen Lautstärke, Tonhöhe und Dauer als unterschiedlich zu empfinden (Terminology, 1973). Es ist also ein mehrdimensionales Attribut, dessen Einzelbestandteile jedoch nicht eindeutig definiert sind. In Martin (1999) wird daher Timbre sogar als schwammiger Begriff bezeichnet, der nicht weiter verwendet

werden sollte. Eng verbunden ist der Begriff Klangfarbe mit dem unterschiedlichen Klang verschiedener Musikinstrumente. Da die Identifikation von Musikinstrumenten deutlich klarer zu definieren ist, wird deren Erkennung häufig als Ersatz für die Schätzung der Klangfarbe verwendet (Emiroglu und Kollmeier, 2008). Auch in dieser Arbeit wird die Instrumentenerkennung verwendet, womit die Zielvariable (Musikinstrument) direkt aus der MIDI-Information herausgelesen werden kann.

Für diese Arbeit werden die folgenden vereinfachenden Annahmen getroffen: Der Fokus des Verfahrens ist klassische Kammermusik, die aus einer fest definierten einstimmigen Melodie- und beliebig vielen Begleitstimmen besteht. Für diese Art der Musik wird im Rahmen des Forschungsprojekts, in dem diese Arbeit entstanden ist, in Kooperation mit einem Projektpartner[2] ein Hörgerätealgorithmus zur Verbesserung der Musikwahrnehmung entwickelt (siehe Kapitel 8). Weiterhin wird davon ausgegangen, dass nur die Erkennung der Eigenschaften der Melodiestimme von Interesse ist, so dass die Begleitung nur als störendes Rauschen betrachtet wird. Verfahren, die nur die Eigenschaften der Melodiestimme extrahieren, werden auch als predominante Verfahren bezeichnet. Es werden somit in dieser Arbeit drei Musikerkennungsaufgaben untersucht: (1) predominante Toneinsatzzeiterkennung, (2) predominante Tonhöhenerkennung und (3) predominante Instrumentenerkennung. Für all diese Aufgaben gibt es bereits umfangreiche Forschungsarbeiten und Standardmethoden, die jedoch in der Regel die Eigenschaften eines Musiksignals auf Basis der eindimensionalen akustischen Wellenform schätzen und nicht auf Basis der mehrdimensionalen Ohrmodellausgabe. Zudem sind für alle Aufgaben die predominanten Varianten erst wenig erforscht.

Ein wichtiger Aspekt an vielen Stellen dieser Arbeit ist der benötigte Rechenzeitbedarf. Dabei muss zwischen drei Perspektiven unterschieden werden:

1. die Rechenzeit für den Hörgerätealgorithmus,
2. die Rechenzeit für die Optimierung des Algorithmus und
3. die Rechenzeit für die Experimente dieser Arbeit.

Der Rechenzeitbedarf des Hörgerätealgorithmus ist durch Realzeitanforderungen und die verwendete Hardware des Hörgeräts hart restringiert. Aber auch eine weitere Reduzierung kann aus Energieeffizienzgründen wichtig sein. Gegenstand dieser Arbeit ist jedoch ausschließlich die Qualitätsverbesserung von Musiksignalen, ohne bereits Rechenzeitaspekte des Hörgeräts zu berücksichtigen. Die Rechenzeit des Hörgerätealgorithmus wird

[2] Institut für Kommunikationsakustik der Ruhr-Universität Bochum

daher nicht weiter betrachtet. Die Optimierung und Evaluation des Hörgerätealgorithmus geschieht hingegen einmalig vor der Benutzung des Hörgeräts und darf insofern länger dauern. Allerdings ist hier zu beachten, dass die Optimierung mehrmals – prinzipiell für jede individuelle Hörschädigung – durchgeführt werden muss, weshalb das Limit der Rechenzeit bei wenigen Tagen liegen sollte. Die Experimente dieser Arbeit wurden auf einem Linux-HPC-Cluster-System[3] durchgeführt. Hier darf die Rechenzeit natürlich höher sein, muss jedoch noch praktikabel bleiben.

Der Aufbau der Arbeit gliedert sich wie folgt. In Kapitel 2 werden Grundlagen beschrieben, die für das Verständnis der weiteren Kapitel wichtig sind. Dazu zählen das verwendete Ohrmodell, musikalische und psychoakustische Grundlagen, Klassifikationsverfahren, Merkmalsselektion sowie modellbasierte Optimierung, die sowohl für die Hörgeräteoptimierung als auch für die Einsatzzeiterkennung benötigt wird. In den folgenden drei Kapiteln werden die drei Musikerkennungsaufgaben detailliert erläutert, wobei jeweils der aktuelle Stand der Forschung und die daraus abgeleiteten Ohrmodellverfahren beschrieben werden. In Kapitel 3 ist dies die Einsatzzeiterkennung, in Kapitel 4 die Tonhöhenerkennung und in Kapitel 5 die Instrumentenerkennung. Anschließend wird in Kapitel 6 ein Versuchsdesign erstellt, wodurch Datensätze für umfassende Vergleichsexperimente der Verfahren definiert werden. In diesem Kapitel wird auch der Aufbau dieser Experimente und die verwendete Software beschrieben. Die Ergebnisse dieser durchgeführten Experimente werden in Kapitel 7 vorgestellt. Sie beinhalten einen Vergleich aller untersuchten Ohrmodellverfahren zu Standardverfahren (ohne Ohrmodell) und zudem auch Ergebnisse für drei beispielhafte Hörschädigungen. In Kapitel 8 wird das entwickelte Bewertungsverfahren praktisch eingesetzt, um einen Hörgerätealgorithmus für eine Hörschädigung zu optimieren. An dieser Stelle wird auch umfangreich diskutiert, wie das Optimierungsverfahren effizienter werden kann. Zum Schluss folgt in Kapitel 9 eine Zusammenfassung der Arbeit, wobei für jeden relevanten Teilaspekt die nächsten sinnvollen Schritte und potentielle Verbesserungen erörtert werden.

[3] `http://lidong.itmc.tu-dortmund.de/ldw/index.php?title=System_overview&oldid=259`

2 Grundlagen

2.1 Ohrmodell

Der Hörvorgang des Menschen und anderer Säugetiere besteht aus mehreren Stufen, die im Ohr und verschiedenen Bereichen des Gehirns stattfinden. Während die höheren Verarbeitungsstufen im Gehirn schwierig zu beobachten sind, ist die erste Stufe im Ohr, die sogenannte auditorische Peripherie, weit besser erforscht. In dieser Stufe werden akustische Schallwellen in der Luft in Aktionspotentiale der Hörnervenfasern transformiert, was durch Computermodelle simuliert werden kann (Ohrmodell). In dieser Arbeit wird das anerkannte und umfangreich analysierte Ohrmodell von Meddis verwendet (Meddis, 2006), dessen aktuellste Version in Panda u. a. (2014) beschrieben ist.

Die auditorische Peripherie besteht aus dem Außen-, dem Mittel- und dem Innenohr (siehe Abbildung 2.1). Die Hauptaufgabe des Außenohrs ist es Schallwellen zu bündeln und diese weiter in das Ohr zu leiten. Am Ende des Außenohrs befindet sich das Trommelfell, das abhängig von der akustischen Eingabe vibriert. Diese Vibration wird weiter an den Steigbügel (lateinisch: *Stapes*), ein Gehörknöchelchen im Mittelohr, geleitet, von wo sie schließlich in der *Cochlea* (deutsch: Hörschnecke), ein Teil des Innenohrs, ankommt. Innerhalb der *Cochlea* befindet sich die Basilarmembran (BM), die abhängig vom Frequenzgehalt des Eingangssignals an verschiedenen Stellen unterschiedlich stark schwingt (siehe Abbildung 2.2). Auf dieser Membran befinden sich innere und äußere Haarzellen. Die inneren Haarzellen werden durch die Schwingung der Basilarmembran angeregt und erzeugen dadurch neuronale Aktivität, auch als Spike Emissionen bezeichnet, die vom Hörnerv (lateinisch: *Nervus Cochlearis Acusticus*) zur Weiterverarbeitung an das Gehirn gesendet wird. Die äußeren Haarzellen dienen zur bewussten Verstärkung bestimmter Frequenzbereiche und erlauben somit die Fixierung auf ein bestimmtes Geräusch. Das auditorische System des Menschen besteht insgesamt aus mehreren Tausend Haarzellen und daran angeschlossenen auditorischen Nervenfasern, wobei jede einem bestimmten

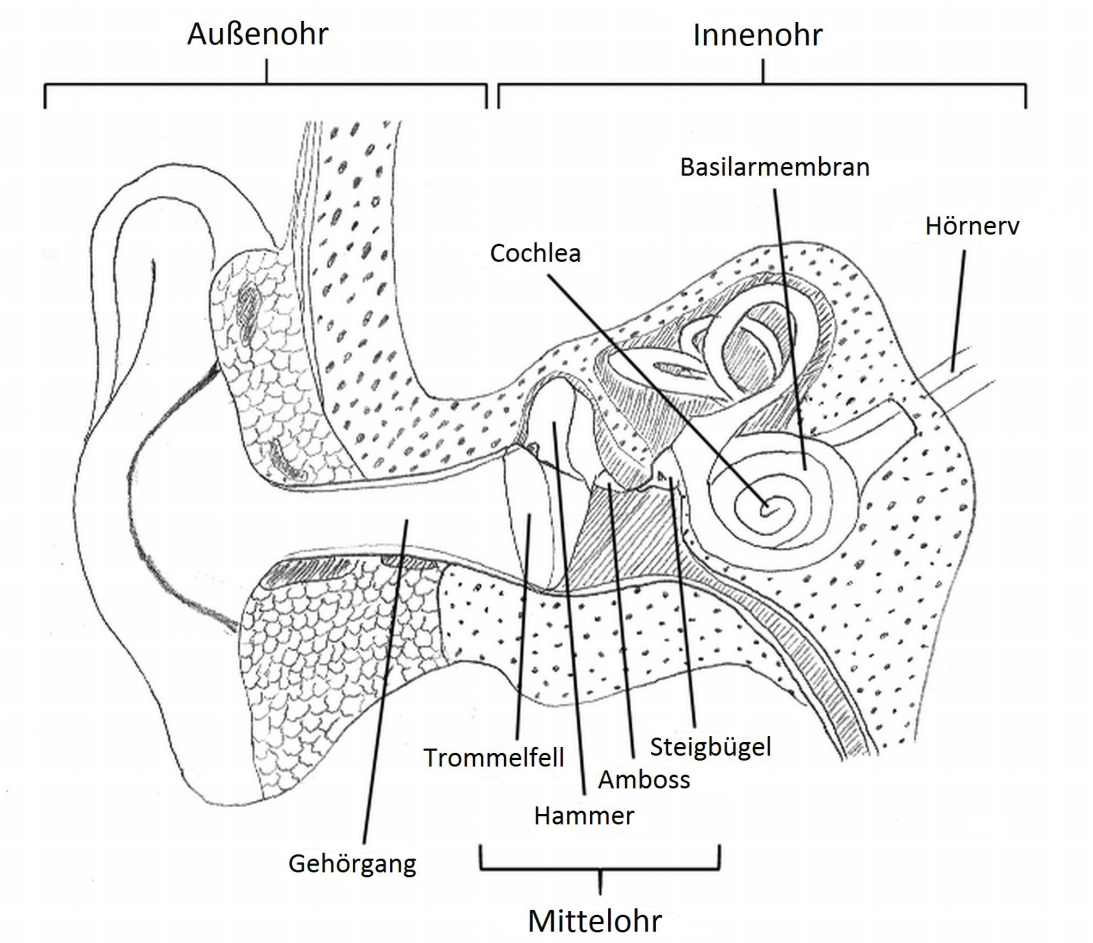

Abbildung 2.1: Modell des menschlichen Ohrs.

Frequenzbereich, die sich jedoch überlappen, zugeordnet werden kann. In den üblichen Simulationsmodellen wird dies jedoch durch eine deutlich kleinere Anzahl vereinfacht.

Das Modell von Meddis ist eine Aneinanderreihung verschiedener Module, welche die beschriebenen Schritte durch mathematische Formeln nachbilden. Der grundlegende Aufbau des Modells ist in Abbildung 2.3 veranschaulicht. Ein Eingangssignal $x[t]$ wird durch vier Module – (1) *Mittelohr*, (2) *Basilarmembran*, (3) *Innere Haarzellen* und (4) *Hörnerv-Synapse* – in Spikewahrscheinlichkeiten $p[t, k]$ (Aktivitätswahrscheinlichkeiten der Hörnervenfasern) transformiert. Dabei bezeichnet t die Zeit und k die simulierte Hörnervenfaser. Da man das Modell signaltheoretisch auch als eine erweiterte Filterbank ansehen kann, wird im Folgenden eine simulierte Hörnervenfaser auch als Kanal bezeichnet.

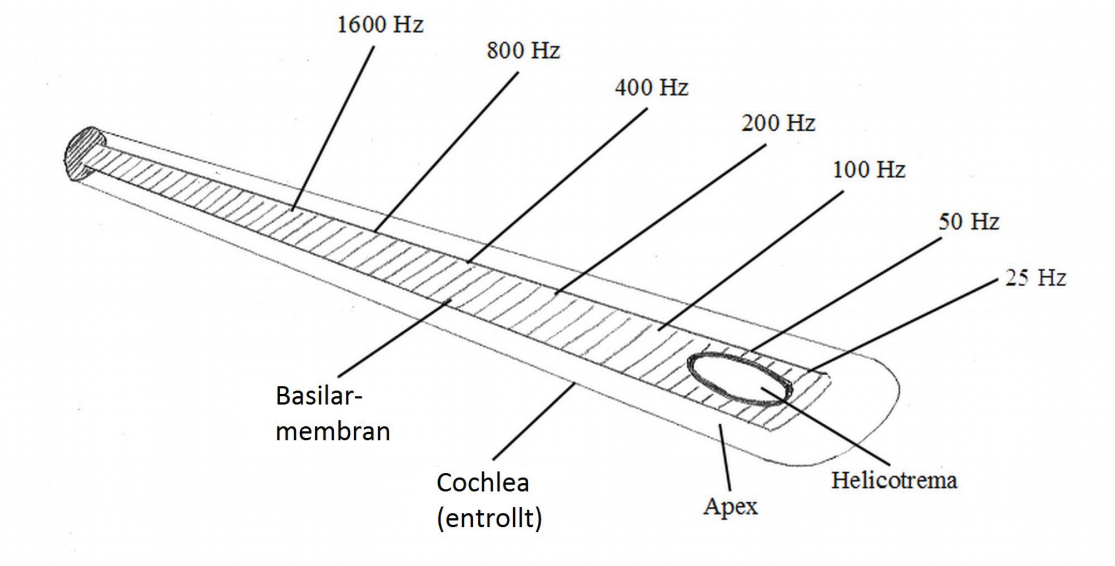

Abbildung 2.2: Frequenzaufteilung der Basilarmembran. An der Spitze der Cochlea (lateinisch: *Apex*) befindet sich das Schneckenloch (lateinisch: *Helicotrema*), durch das der transformierte Schall die Basilarmembran erreicht. In der Nähe des Apex wird die Basilarmembran von tiefen Frequenzen angeregt und an ihrer Basis von hohen Frequenzen.

Wie im realen menschlichen Hörsystem besitzt auch im Simulationsmodell jeder Kanal eine individuelle charakteristische Frequenz, die auch als *Best Frequenz* (BF) bezeichnet wird. Diese Frequenz definiert, auf welche Frequenz eines Eingangssignals der Kanal am stärksten reagiert, wobei jeder Kanal prinzipiell wie ein Bandpassfilter arbeitet. Das heißt die Stärke der Reaktion auf eine bestimmte Frequenz ist abhängig von ihrem Abstand zur *Best Frequenz*. Im Modell von Meddis, das mit den Grundeinstellungen benutzt wird, gibt es insgesamt 41 Kanäle, deren *Best Frequenzen* zwischen 100 hz und 6 kHz liegen.[1]

Die einzelnen Transformationsschritte des Ohrmodells (vergleiche Abbildung 2.3) werden in den folgenden Abschnitten detailliert erläutert. In Abbildung 2.4 sind die Ausgaben dieser Schritte zudem an einem Beispiel veranschaulicht.

[1] Für einige Vorversuche, die in dieser Arbeit beschrieben werden, wurden auch ältere Versionen des Modells verwendet, die eine andere Anzahl an Kanälen mit anderen Frequenzbereichen verwenden. In diesen Fällen wird explizit darauf hingewiesen, ansonsten ist immer die neuste Version des Modells gemeint.

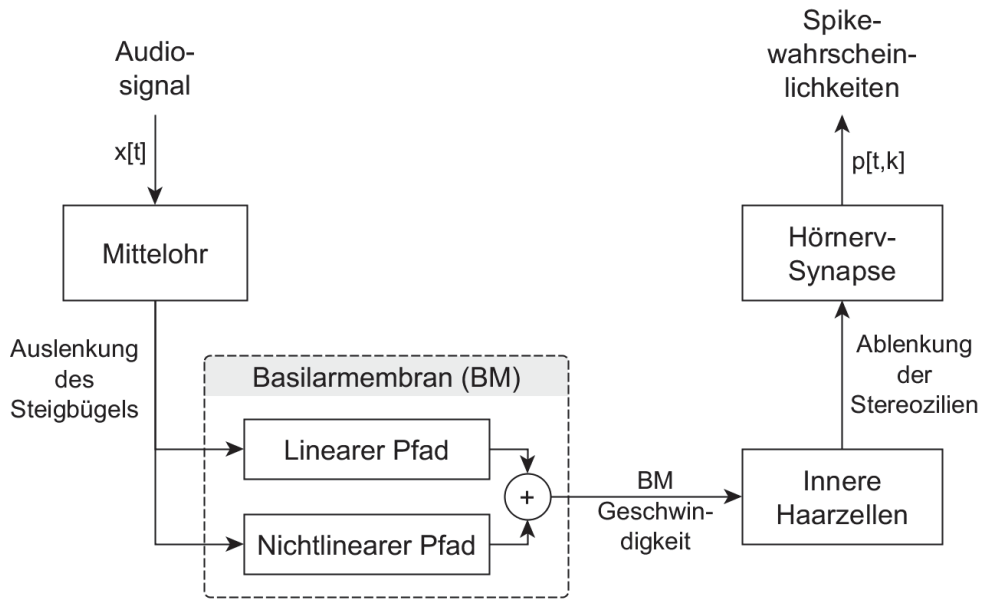

Abbildung 2.3: Blockdiagramm des Ohrmodells von Meddis.

2.1.1 Außen- und Mittelohr

Die Funktion des Außen- und Innenohrs wird in dem Modell etwas vereinfacht, indem im ersten Schritt die akustischen Schwingungen in der Luft direkt in die dadurch entstehende Geschwindigkeit des Steigbügels umgerechnet wird. Die Spitzengeschwindigkeit des Steigbügels ist dabei proportional zu dem Spitzendruck am Trommelfell bis etwa 130 db SPL (*Sound Pressure Level* = Schalldruckpegel). Im Modell wird dies durch eine Kombination von speziellen Filtern erreicht (Sumner u. a., 2003). In der neuesten Version des Modells, wurde die Geschwindigkeit des Steigbügels durch die räumliche Auslenkung ersetzt, wodurch die Anzahl der Parameter für das folgende Modul der Basilarmembran reduziert werden kann, ohne dass das Modell an Genauigkeit verliert (Panda u. a., 2014). Eine beispielhafte Ausgabe des Mittelohrs ist im zweiten Plot von Abbildung 2.4 zu sehen.

2.1.2 Basilarmembran

Im nächsten Schritt wird die Auslenkung des Steigbügels (bzw. im älteren Modell dessen Geschwindigkeit) in die Geschwindigkeit der Basilarmembran an verschiedenen Stellen

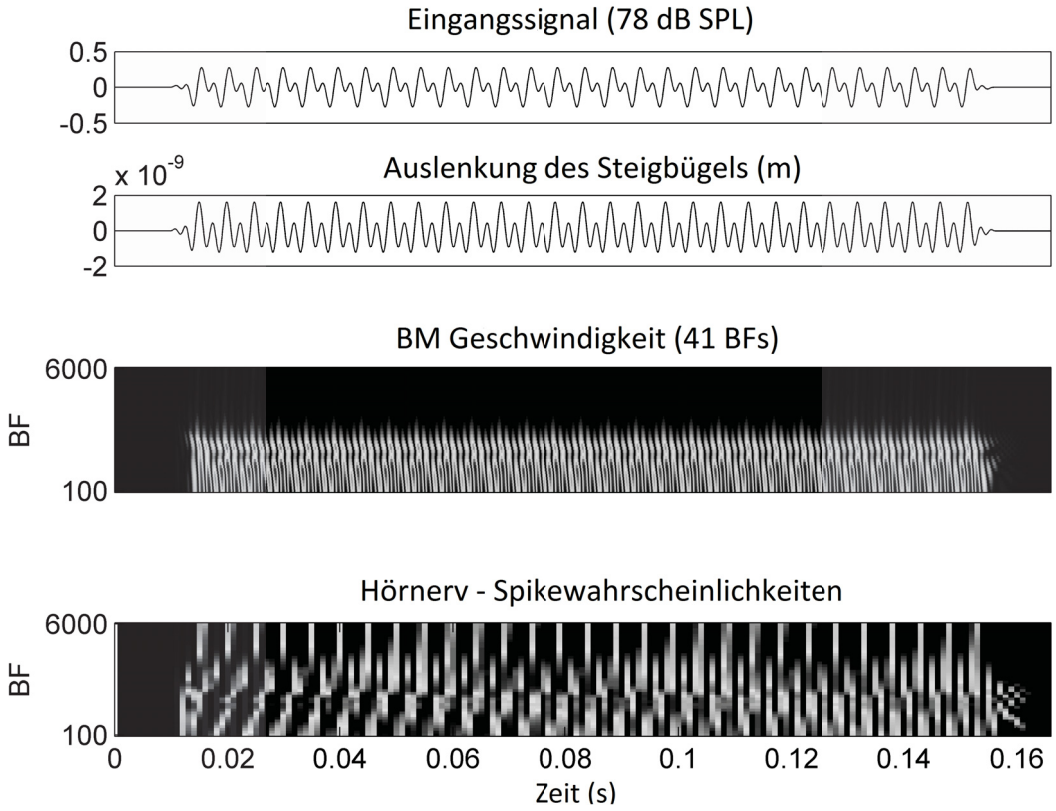

Abbildung 2.4: Beispielausgabe des Ohrmodell von Meddis: (1) Akustisches Signal (200 Hz + 400 Hz), (2) Ausgabe des Mittelohrs (Auslenkung des Steigbügels), (3) Ausgabe der Basilarmembran (BM) in Abhängigkeit von den *Best Frequenzen* (BF) der Kanäle, (4) Ausgabe der Hörnervenfasern in Abhängigkeit von den BFs.

umgewandelt, was durch eine Menge von überlappenden Filtern (Kanäle) implementiert wird. Jeder Kanal entspricht dabei einer bestimmten Stelle auf der Basilarmembran und reagiert am stärksten auf Frequenzen die nahe seiner spezifischen *Best Frequenz* (BF) sind. Die *Best Frequenzen* der Kanäle sind logarithmisch verteilt mit äquidistanten Abständen (vergleiche auch Abbildung 2.2). Die *Best Frequenz* eines Kanals $f_{BF}[k]$ ist definiert durch

$$f_{BF}[k] = 10^{\log_{10}(BF_{lowest}) + \frac{k-1}{K-1}[\log_{10}(BF_{highest}) - \log_{10}(BF_{lowest})]}, \quad k = 1, \ldots, K. \tag{2.1}$$

Dabei bezeichnet k die Kanalnummer, K die Anzahl der Kanäle, BF_{lowest} die niedrigste und $BF_{highest}$ die höchste *Best Frequenz*. Für die neueste Version des Modells gilt somit

$K = 41$, $BF_{lowest} = 100$ und $BF_{highest} = 6000$. In früheren Ohrmodellen (z.B. Slaney u. a., 1993) wurde der gesamte Prozess der Basilarmembran durch eine *Gammatone*-Filterbank modelliert, die aus überlappenden, linearen und symmetrischen Bandpassfiltern besteht. Ein *Gammatone*-Filter ist definiert durch seine Impulsantwort (Reaktion bzw. Ausgangssignal des Filters auf einen kurzen Eingangsimpuls):

$$g(t) = at^{n-1} \exp(-2\pi bt) \cos(2\pi f_{BF} t + \phi) \quad \forall t > 0. \tag{2.2}$$

Dabei ist t die Zeit, a ein Verstärkungsfaktor, n die Filterordnung, f_{BF} die *Best Frequenz*, b die Bandweite und ϕ die Phase der Feinstruktur der Impulsantwort (Patterson u. a., 1992). Die Bandweite ist proportional zur *Best Frequenz* und wird definiert durch

$$b = 24.7 \cdot (^{4.37 \cdot f_{BF}}/_{1000} + 1) \cdot BW_C, \tag{2.3}$$

mit der Bandweitenkorrektur BW_C, die abhängig von der Filterordnung ist. Für $n = 4$ ist $BW_C = 1.019$ (Patterson u. a., 1992).

Für eine realistischere Wiedergabe der menschlichen Wahrnehmung müssen die Filter jedoch nichtlinear und asymmetrisch sein. In jedem Kanal sollten höhere Frequenzen stärker abgeschwächt werden als tiefere und schwächere Schallpegel sollten mehr verstärkt werden. In Meddis Modell werden diese gewünschten Effekte durch eine *Dual-Resonance-Non-Linear* (DRNL) Filterbank modelliert, die einer nichtlinearen Erweiterung der *Gammatone*-Filterbank mit einem zusätzlichen Tiefpassfilter entspricht. Durch diese Änderung konnte gezeigt werden, dass psychoakustische Phänomene, wie Maskierungseffekte (z.B. Verdeckung eines Tons durch einen anderen), in dem Modell berücksichtigt sind.

Die DRNL Filterbank besteht aus zwei asymmetrischen Bandpassfiltern die parallel arbeiten: Einen linearen und einen nichtlinearen Pfad (vergleiche Abbildung 2.3). In der Realität sind außerdem die *Best Frequenzen* und die Bandweiten abhängig von der Lautstärke. Stärkere Pegel führen zu einer Senkung der *Best Frequenz* und einer Erweiterung der Bandweiten. In der DRNL Filterbank wird dies durch tiefere *Best Frequenzen* und höhere Bandweiten im nichtlinearen Pfad erreicht, denn bei höheren Lautstärkepegeln steigt der Einfluss dieses Pfades und der Einfluss des linearen Pfades sinkt. Die einzelnen Schritte beider Pfade sind detailliert in Abbildung 2.5 skizziert. Der nichtlineare Pfad besteht aus einer Kombination aus einem *Gammatone*-Filter, einer nichtlinearen Kompressionsfunktion (*Broken-Stick*), einem weiteren *Gammatone*-Filter und einem Tiefpassfilter (*Butterworth*-Filter). Der lineare Pfad besteht aus einer Kombination aus

einer Verstärkungsfunktion, einem *Gammatone*-Filter und einem Tiefpassfilter (*Butterworth*-Filter). Am Ende werden beide Pfade durch Summierung wieder zusammengefügt. Für das vorher bereits verwendete Tonbeispiel kann man die entsprechende Ausgabe der Basilarmembran im dritten Plot von Abbildung 2.4 sehen.

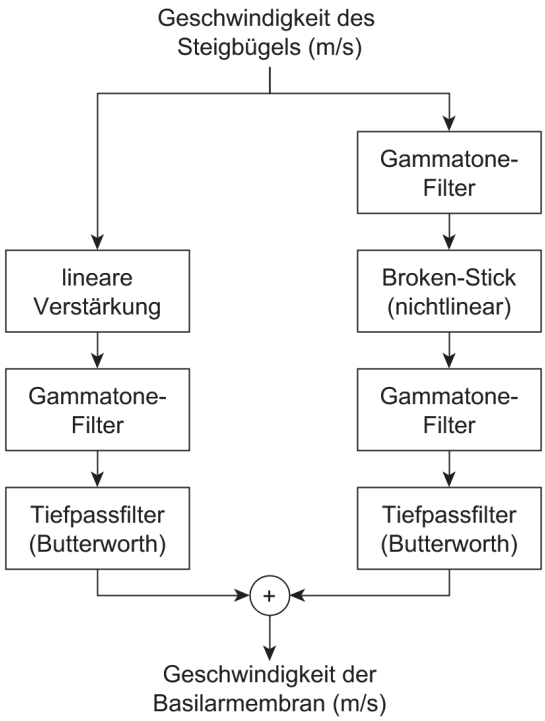

Abbildung 2.5: Aufbau der DRNL-Filterbank (Lopez-Poveda und Meddis, 2001).

2.1.3 Innere Haarzellen

Auf der Basilarmembran befinden sich innere Haarzellen, an deren Spitze sich ein Bündel von jeweils 20 bis 300 winzigen Wimpern (Stereozilien) befindet (Roederer und Mayer, 1999). Innerhalb dieser Zellen wird die mechanische Schwingung der Basilarmembran in elektrisches Potential umgewandelt, das anschließend zu der Ausschüttung von Neurotransmitter führt. Zunächst wird die zeitabhängige Geschwindigkeit der Basilarmembran $v(t)$ in eine zeitabhängige Verschiebung der Stereozilien $u(t)$ transformiert, definiert durch

$$\tau_c \frac{du(t)}{dt} + u(t) = \tau_c C_{cilia} v(t), \tag{2.4}$$

wobei τ_c eine Zeitkonstante und C_{cilia} ein Verstärkungsfaktor ist. Die Ablenkung der Stereozilien stärkt bzw. hindert den Eingangsfluss von Ionen in die Zelle und bestimmt damit ihr elektrisches Potential. Dies erzeugt eine asymmetrische Schaltung des Übertragungskanals, denn positive Ablenkungen der Stereozilien werden durch den Inonenfluss weitergeleitet, aber negative Ablenkungen vollständig geblockt.

2.1.4 Hörnerv-Synapse

Im nächsten Schritt wird die Verschiebung der Stereozilien durch eine Calcium-kontrollierte Transmitterausschüttungsfunktion in neuronale Aktivitätswahrscheinlichkeiten $p[t, k]$ transformiert, wobei t die Zeit und k die Kanalnummer bezeichnet. Transmittersubstanz wird ausgeschüttet, wobei deren Rate vom elektrischen Potential der Haarzelle und von der Menge an vorhandener Transmittersubstanz abhängt, die in Vesikeln (kleine Bläschen) im synaptischen Spalt der Haarzelle gespeichert wird. Eine Steigerung des elektrischen Potentials steigert auch die Wahrscheinlichkeit einer Transmitterausschüttung. Eine Serie von Ausschüttungen senkt allerdings das Reservoir an Transmitter und reduziert dadurch die Wahrscheinlichkeit von weiteren Ausschüttungen bis der lokale Speicher wieder aufgefüllt ist.

Tatsächlich werden Transmitterausschüttungen nur indirekt durch die elektrische Spannung $V(t)$ gesteuert, denn diese steuert lediglich den Calciumfluss in die Zelle, was dann die Transmitterausgabe fördert. Die Steigerung der Calciumströmung $I_{Ca}(t)$ wird gesteuert durch

$$I_{Ca}(t) = G_{Ca}^{max} m_{I_{Ca}}^3(t)(V(t) - E_{Ca}). \tag{2.5}$$

Dabei ist G_{Ca}^{max} die Calciumleitfähigkeit in der Umgebung der Synapse, wenn alle Übertragungskanäle offen sind, $m_{I_{Ca}}^3(t)$ ist der Anteil von offenen Calciumkanälen und E_{Ca} ist das Umkehrpotential von Calcium (das Potential, bei dem die Calciumströmung neutral ist). Während in vielen anderen Ohrmodellen die vereinfachende Annahme gemacht wird, dass der Beitrag des Transmitterreservoirs im synaptischen Spalt proportional zu der Intensität des Stimulus ist, wird im Modell von Meddis die Transmitterausschüttung realistischer durch eine Aneinanderreihung mehrerer Transmitterreservoire modelliert. Für das Beispiel in Abbildung 2.4 sind die neuronale Aktivitätswahrscheinlichkeiten $p[t, k]$ – auch als Spikewahrscheinlichkeiten oder Spikefeuerraten bezeichnet – im untersten Plot zu sehen.

2.1.5 Hörnervaktivität

Im letzten Schritt des Simulationsmodells werden Aktivitätswahrscheinlichkeiten in binäre Ausschüttungsereignisse (Spikes) umgewandelt. Dies ist ein stochastischer Prozess, wobei die Wahrscheinlichkeit eines Spikes abhängig von der Ausschüttungswahrscheinlichkeit und der zeitlichen Entfernung des letzten generierten Spikes ist. Für jeden Kanal werden n Hörnervenfasern durch eine n-fache Wiederholung dieses Prozesses simuliert (typischerweise $n \in [10, 100]$). Für die Anwendungen in dieser Arbeit sind allerdings lediglich die Aktivitätswahrscheinlichkeiten von Interesse, und die Umwandlung in binäre Ereignisse, die sehr rechenintensiv ist, kann vernachlässigt werden.

2.1.6 Ohrmodelle mit simulierten Hörschädigungen: *Hearing Dummies*

Es werden drei beispielhafte Hörmodelle mit simulierten Hörschädigungen – im folgenden auch als *Hearing Dummies* bezeichnet – betrachtet, die in Meddis u. a. (2010a) und Panda u. a. (2014) beschrieben sind. Es sind modifizierte Varianten des Ohrmodells von Meddis, die durch Änderungen möglichst weniger Parameterwerte an reale Hörschädigungen angepasst sind. In der ursprünglichen Fassung wurden für die *Hearing Dummies* allerdings nur Kanäle mit *Best Frequenzen* über 250 Hz untersucht. Damit diese mit dem oben beschriebenen Modell ohne Hörschädigung besser vergleichbar sind und damit auch für musikalische Töne mit Grundfrequenzen unter 250 Hz sinnvolle Ergebnisse erzielt werden können, werden die *Dummies* entsprechend der *Best Frequenzen* des Modells ohne Hörschädigung modifiziert, was dazu führt, dass neun Kanäle mit *Best Frequenzen* zwischen 100 und 250 Hz hinzugefügt werden.

Der erste *Hearing Dummy* simuliert einen starken Mittel- und Hochfrequenzhörverlust. Der ursprünglich dafür vorgeschlagene *Hearing Dummy* behält dabei nur den tiefsten Kanal mit der *Best Frequenz* von 250 Hz, bei dem zusätzlich noch der nichtlineare Pfad der DRNL Filterbank entfernt wird. In der hier verwendeten modifizierten Version des *Dummies* werden die ersten 10 Kanäle behalten (alle mit *Best Frequenzen* bis maximal 250 Hz), wobei der nichtlineare Pfad all dieser Kanäle entfernt wird. Der zweite *Hearing Dummy* simuliert eine Mittelfrequenzschwerhörigkeit mit einer klaren Störung in der Frequenzregion zwischen 1 und 2 kHz. Um dies zu erreichen, werden in der modifizierten Version des *Dummies* 16 Kanäle entfernt (Kanäle 17 bis 32). Der dritte *Hearing Dummy* simuliert eine steile Hochtonschwerhörigkeit, was durch den Wegfall aller Kanäle mit *Best Frequenzen* über 1750 Hz, was den letzten zwölf Kanälen entspricht, erreicht wird.

2.2 Musikalische und Psychoakustische Grundlagen

Die menschliche Wahrnehmung eines Tons ist definiert durch vier Dimensionen: die Tonhöhe, die Lautheit, die Tondauer und die Klangfarbe. Natürliche Töne bestehen im Wesentlichen aus sich überlagernden Schwingungen und werden daher auch als komplexe Töne bezeichnet. Dagegen wird ein Ton, der nur aus einer einzelnen Schwingung besteht, als reiner Ton bezeichnet und ist durch

$$x[t] = A\sin(2\pi ft + \phi), \quad t = 1, \ldots, M \tag{2.6}$$

definiert, wobei t den Zeitindex, A die Amplitude, f die Frequenz, ϕ die Phase und M die Anzahl an Abtastwerten (Tondauer) bezeichnet. Reine Töne kommen nicht natürlich vor und klingen stumpf und unangenehm, können aber künstlich erzeugt werden (Roederer und Mayer, 1999). Auf Grund der komplizierten Struktur komplexerer Töne basieren viele psychoakustische Untersuchungen, wie beispielsweise die Lautstärkewahrnehmung (Lautheit), lediglich auf reinen Tönen.

Die Tonhöhe eines reinen Tons ist direkt abhängig von der Frequenz f. Die Frequenzbestandteile eines Signals können mit Hilfe der Fourier-Transformation ermittelt werden. Bei realen Anwendungen liegen allerdings Signale üblicherweise in einer zeitdiskreten Form vor (z.B. 44100 Abtastwerte pro Sekunde bei CD-Qualität). In diesem Fall verwendet man die diskrete Fourier Transformation (DFT), die durch

$$X[\mu] = \sum_{t=0}^{M-1} x[t] \exp\left(-i\frac{2\pi\mu t}{M}\right), \quad \mu = 0, \ldots, M-1 \tag{2.7}$$

definiert ist. Auf Grund der Diskretisierung hat man, im Unterschied zur Fourier Transformation, bei der DFT keine perfekte Frequenzauflösung, sondern der Frequenzbereich wird in äquidistante Frequenzbereiche (Frequenzbänder, englisch: *frequency bins*) aufgeteilt, deren Mittenfrequenzen auch als Frequenzlinien bezeichnet werden. Die Frequenzkoeffizienten werden mit $X[\mu]$ bezeichnet und definieren das Spektrum des μ-ten Frequenzbandes bzw. der μ-ten Frequenzlinie. $X[\mu]$ ist eine komplexe Zahl, deren Imaginärteil die Phase beschreibt. Allerdings ist man oft – wie auch in allen Anwendungen in dieser Arbeit – lediglich an den Intensitäten der Frequenzen interessiert. Diese sind durch den Betrag der DFT-Koeffizienten $|X[\mu]|$ definiert, der als spektrale Amplitude bezeichnet wird. Bei einem Signal $x[t]$, das nur aus einem einzigen reinen Ton mit der Frequenz f besteht (siehe Gleichung 2.6), ist $|X[\mu]|$ maximal für die Frequenzlinie μ, die der Frequenz f am

nächsten liegt. Eine andere Möglichkeit diese Frequenz zu bestimmen, ist die Betrachtung der Periode T im Zeitbereich, denn es gilt $f = \frac{1}{T}$. Die Periode kann mit Hilfe der Autokorrelationsfunktion (ACF) ermittelt werden (siehe Kapitel 4).

Die Lautstärke eines reinen Tons wird durch die Amplitude A bestimmt, aber die Lautheit (also die wahrgenommene Lautstärke) ist auch abhängig von der Tonhöhe und der Tondauer. Die Phase ϕ hat bei reinen Tönen keinen Einfluss auf die Wahrnehmung der musikalischen Dimensionen, sie hilft allerdings bei der Lokalisierung und Trennung verschiedener Schallquellen. Bei der Überlagerung verschiedener Töne kann sie aber auch einen Einfluss auf die Lautheit haben. Ein Beispiel dafür ist die destruktive Interferenz, bei der zwei reine Töne mit gleicher Frequenz und gleicher Amplitude sich durch eine Phasenverschiebung um 180° vollkommen aufheben, so dass die Töne nicht mehr zu hören sind (Roederer und Mayer, 1999).

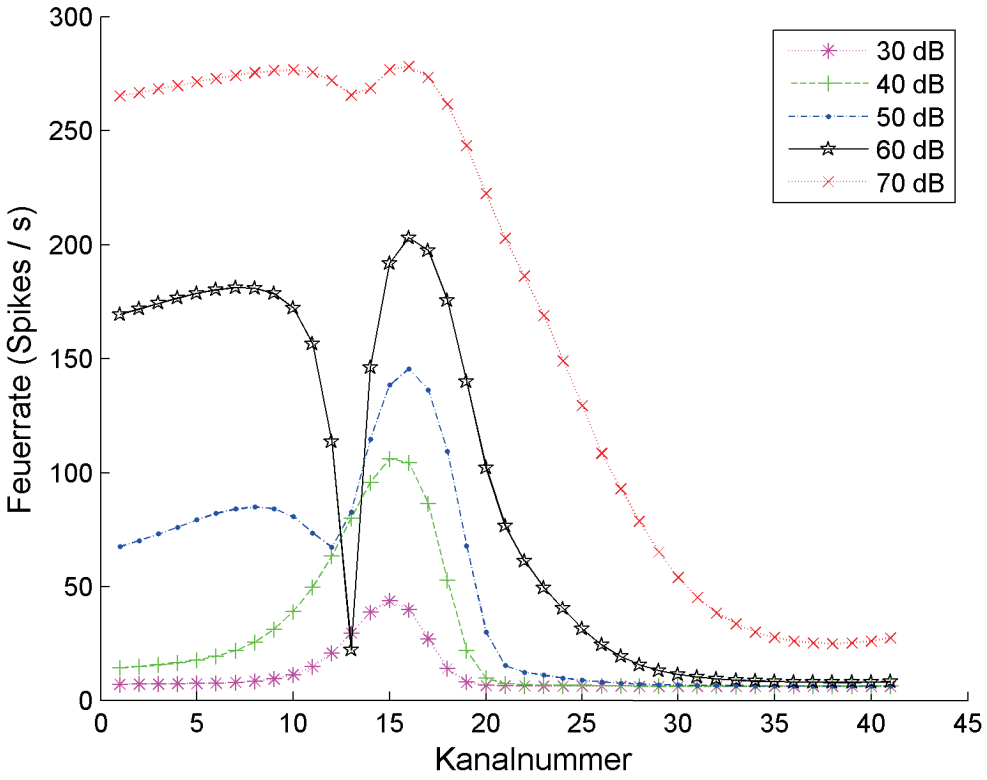

Abbildung 2.6: Mittlere Feuerraten der Kanäle bei einem Ton von 440 Hz für unterschiedliche Lautstärken (in dB SPL).

Im menschlichen Ohr wirkt sich die Tonhöhe auf die Schwingung der Basilarmembran

und damit auf die Spikefeuerraten der verknüpften Hörnervfasern aus (vergleiche Kapitel 2.1). Dabei wird die Tonhöhe gleich doppelt kodiert: Zum einen legt sie fest, welche Region der Basilarmembran am stärksten schwingt und damit welche Fasern wie stark feuern, und zum anderen bestimmt sie, in welcher Frequenz Spikes gefeuert werden. Dies geschieht nämlich nicht in einer konstanten Rate, sondern Phasen mit hohen und niedrigen Feuerraten wechseln sich ab und für Frequenzen bis mindestens 2 kHz geschieht dies in einer Frequenz, die identisch zur Tonfrequenz ist (*Phaselocking Effekt*). Diese Charakteristik ist natürlich auch im Ohrmodell zu beobachten. Abbildung 2.6 zeigt die mittleren Feuerraten der Kanäle des Ohrmodells für einen reinen Ton mit der Frequenz 440 Hz bei verschiedene Lautstärken. Unabhängig von der Lautstärke ist die Aktivität der Kanäle 15 und 16, die eine *Best Frequenz* von 419 bzw. 464 Hz haben[2], am höchsten. Zudem wird in der Regel bei höheren Lautstärken auch die Aktivität aller Kanäle höher. Es gibt aber auch Ausnahmen, denn vor allem für Kanal 13 ist bei einer Lautstärke von 60 dB SPL die mittlere Feuerrate unerwartet niedrig. Grund dafür ist die Wellenbewegung der Basilarmembran, die sich aus einem linearen und einem nichtlinearen Bestandteil zusammenaddiert (siehe Kapitel 2.1.2). Für beide Bestandteile ergeben sich leichte Verzögerungen, die abhängig vom Abstand der Frequenz des Eingangssignals zu der jeweiligen *Best Frequenz* sind, die für beide Pfade unterschiedlich sind. Für den Beispielton sind bei Kanal 13 dadurch die Phasen annähernd um 180° verschoben, und zudem sind auch die beiden Bestandteile in etwa gleich groß, so dass sie sich ähnlich wie bei der destruktiven Interferenz gegenseitig aufheben (siehe auch Abbildung B.1 in Anhang B). Dagegen ist bei niedrigeren Lautstärken der lineare Pfad dominierend und bei höheren Lautstärken der nichtlineare Pfad. Für die Tonhöhenwahrnehmung wirkt sich dieser Effekt allerdings nicht negativ aus, denn hierfür sind die Kanäle wesentlich, deren *Best Frequenzen* nahe der Tonfrequenz sind (also hier Kanal 15 und 16), und dieser Effekt kann nur bei höheren Frequenzunterschieden auftauchen.

Genauer ist die Frequenz anhand der DFT eines Kanals zu erkennen, wie beispielhaft in Abbildung 2.7 gezeigt. Am deutlichsten ist das Maximum bei 440 Hz in Kanal 15 sichtbar, aber auch in den anderen Kanälen ist die Frequenz 440 Hz maximal. Dies gilt aber natürlich nicht mehr, sobald auch andere Frequenzen im Signal enthalten sind. Weiterhin fällt auf, dass auch bei allen ganzzahligen Vielfachen von 440 Hz lokale Maxima deutlich zu erkennen sind, obwohl diese Frequenzen nicht im Eingangssignal enthalten sind.

Die Töne der meisten Musikinstrumente bestehen im Wesentlichen aus Überlagerungen

[2] für den linearen Pfad, im nichtlinearen Pfad sind die *Best Frequenzen* etwas tiefer (vergleiche Kapitel 2.1.2)

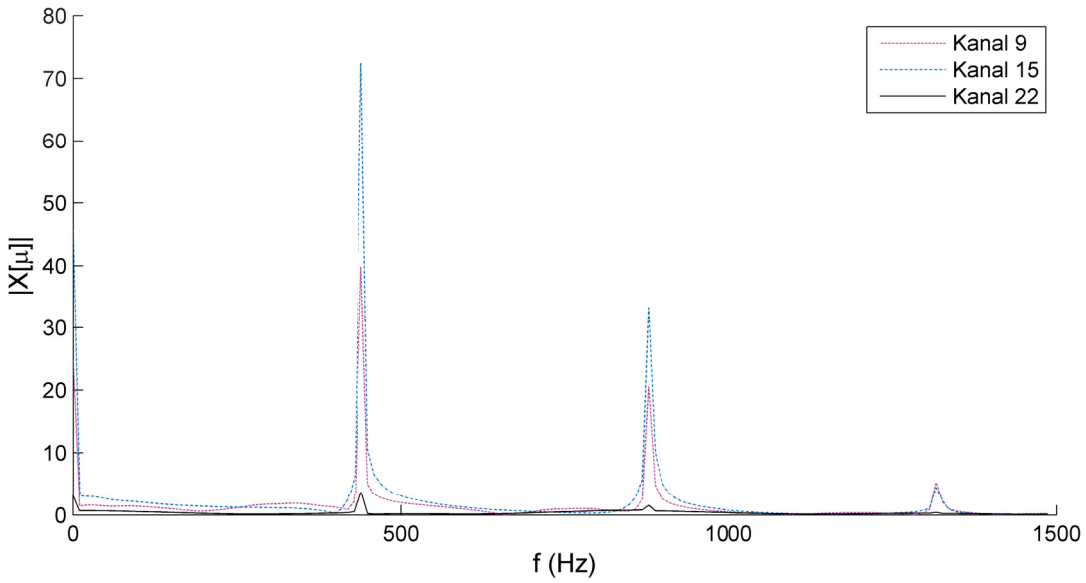

Abbildung 2.7: DFT-Amplituden der Kanäle 9, 15 und 22 bei einem reinen Ton mit einer Frequenz von 440 Hz und einer Lautstärke von 50 dB SPL.

harmonischer Schwingungen. Dies bedeutet, ein solcher Ton besteht aus einem Grundton mit der Grundfrequenz F_0 und Obertönen, deren Frequenzen F_n ganzzahlige Vielfache von F_0 sind. Statt der Aufteilung in Grundton und Obertöne wird auch oft die allgemeinere Bezeichnung Partialton (oder auch Harmonische) verwendet. Dabei muss beachtet werden, dass der erste Partialton dem Grundton und der n-te Partialton dem $(n-1)$-ten Oberton entspricht (für $n > 1$). Ein idealer harmonischer Ton, der nur aus harmonischen Bestandteilen besteht, ist durch

$$x[t] = \sum_{n=1}^{N_{\text{part}}} A_n \sin(2\pi n f t + \phi_n), \quad t = 1, \ldots, M \tag{2.8}$$

definiert, wobei n die Partialtonnummer und N_{part} den maximal betrachteten Partialton bezeichnet. Die Verteilung der Intensitäten auf die Partialtöne ist entscheidend für die wahrgenommene Klangfarbe und kann daher dazu verwendet werden Musikinstrumente zu unterscheiden. Allerdings ist diese Verteilung für unterschiedliche Tonhöhen sehr verschieden, und sie ist zudem auch abhängig von der Lautstärke, der Spielweise und der Raumakustik. Man nimmt an, dass vor allem die ersten sieben bis acht Partialtöne für die Musikinstrumentenerkennung entscheidend sind, da höhere Partialtöne bei der

Wahrnehmung nicht mehr einzeln erfasst werden können, sondern zusammenfallen (Roederer und Mayer, 1999). Neben der Intensitätsverteilung der Partialtöne sind aber auch noch weitere Charakteristika des Tons entscheidend für die Instrumentenerkennung, wie z.B. die Entwicklung des Tonanschlags, unharmonische Bestandteile oder periodische Änderungen der Frequenz (Vibrato) oder der Amplituden (Tremolo).

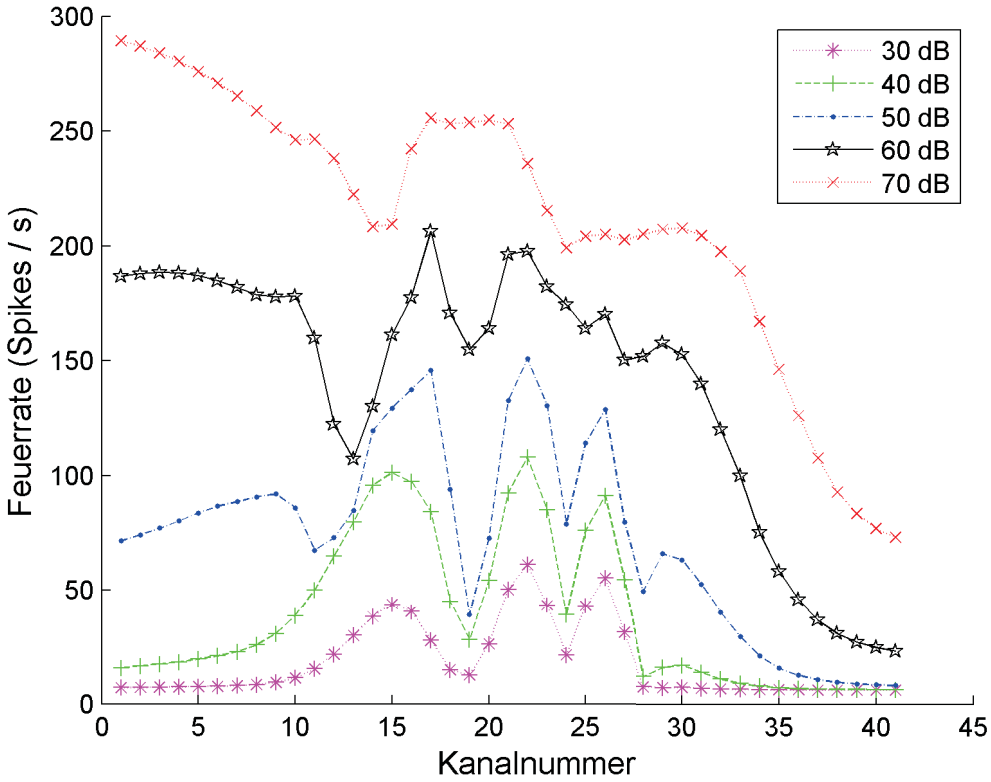

Abbildung 2.8: Mittlere Feuerraten der Kanäle bei einem harmonischen Ton mit einer Grundfrequenz von 440 Hz, der aus den ersten drei Partialtönen mit identischen Intensitäten besteht, für unterschiedliche Lautstärken (in dB SPL).

Abbildung 2.8 zeigt die mittleren Feuerraten der Ohrmodellkanäle für einen harmonischen Ton, der neben der Grundfrequenz von 440 Hz mit identischen Intensitäten auch den zweiten und dritten Partialton enthält. Im Unterschied zu Abbildung 2.6 sieht man hier neben hohen Feuerraten um Kanal 15 auch hohe Feuerraten um die Kanäle 22 und 26, deren *Best Frequenzen* nahe den beiden Partialtonfrequenzen liegen. Bei hohen Lautstärken, bei denen der nichtlineare Pfad der Basilarmembran dominant ist, sind diese Maxima allerdings nicht mehr zu erkennen. In Abbildung 2.9 sind die spektralen Amplituden dieses Tons für drei ausgewählter Kanäle zu sehen. Der wesentliche Unter-

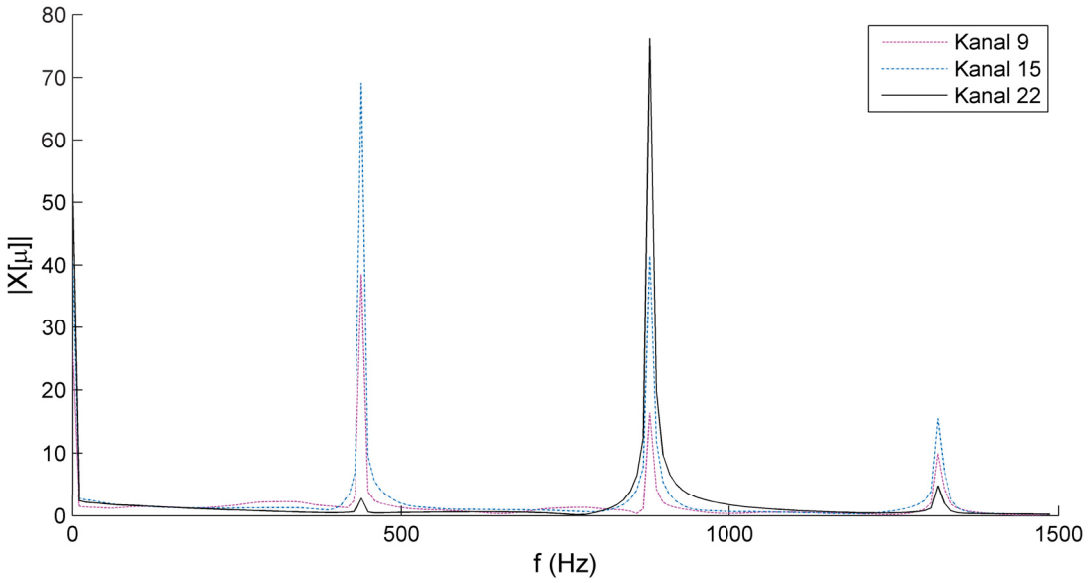

Abbildung 2.9: DFT-Amplitude der Kanäle 9, 15 und 22 bei einem harmonischen Ton mit einer Grundfrequenz von 440 Hz, der aus den ersten drei Partialtönen mit identischen Intensitäten besteht, und einer Lautstärke von 50 dB SPL.

schied zu Abbildung 2.7 ist bei Kanal 22 zu sehen, der jetzt eine starke Intensität bei der Frequenz des zweiten Partialtons (880 Hz) zeigt.

Die Abbildungen 2.10 und 2.11 zeigen, wie sich Klangfarbenunterschiede der Instrumente Cello, Klarinette und Trompete auf die Ohrmodellausgabe für einen Ton mit der Grundfrequenz von 440 Hz auswirken. In Abbildung 2.10, welche die mittleren Feurraten der Kanäle zeigt, unterscheidet sich vor allem die Klarinette von den beiden anderen Instrumenten, denn bei ihr zeigen die Kanäle, deren *Best Frequenzen* den geraden Partialtönen entsprechen, nur schwache Aktivitäten. Dies ist darin begründet, dass der Klang der Klarinette genau diesen Effekt aufweist, und die Intensitäten fast ausschließlich auf die ungeraden Partialtöne verteilt sind. Abbildung 2.11 zeigt die spektralen Amplituden des Kanals 15 in der Umgebung der Grundfrequenz von 440 Hz. Hier ist vor allem auffällig, dass beim Cello die Intensitäten weiter gestreut sind, wofür dessen Vibratoklang verantwortlich ist. Zudem fällt bei der Trompete noch auf, dass das Maximum der Intensität etwas zu hoch liegt, wofür aber lediglich eine nicht exakte Stimmung des Instruments verantwortlich sein dürfte.

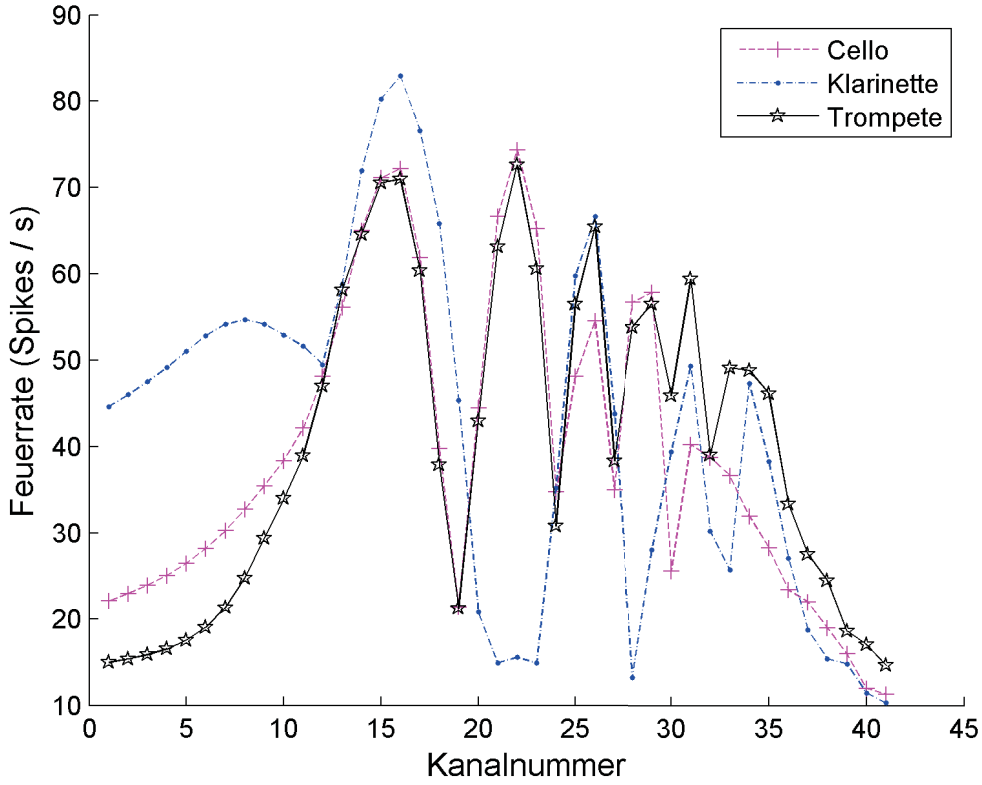

Abbildung 2.10: Mittlere Feuerraten der Kanäle bei realen Tönen (McGill University Master Samples (Eerola und Ferrer, 2008)) mit einer Grundfrequenz von 440 Hz dreier Musikinstrumente.

2.3 Klassifikationsverfahren

Für die Verfahren zur Tonhöhenschätzung und Instrumentenerkennung wird überwachte Klassifikation benötigt. Klassifikation ist formal eine Abbildung $f : \Phi \rightarrow \Psi$, wobei Φ der Eingaberaum ist, der spezielle Charakteristika der zu klassifizierenden Beobachtungen beschreibt, und Ψ ist die Menge der Kategorien bzw. Klassen. Bei den hier vorkommenden Klassifikationsaufgaben ist Φ eine Menge von Merkmalen, die etwas über den spektralen oder zeitlichen Verlauf eines Signals aussagen und Ψ ist entweder eine Menge von Musikinstrumenten oder eine Menge von Frequenzbereichen.

Für die Experimente in dieser Arbeit ist es aus Rechenzeitgründen wichtig, dass das verwendete Klassifikationsverfahren auch ohne eine umfangreiche Parameter-Optimierung verlässliche Resultate erzielt. Andererseits sollen aber aus Vergleichsgründen zumindest

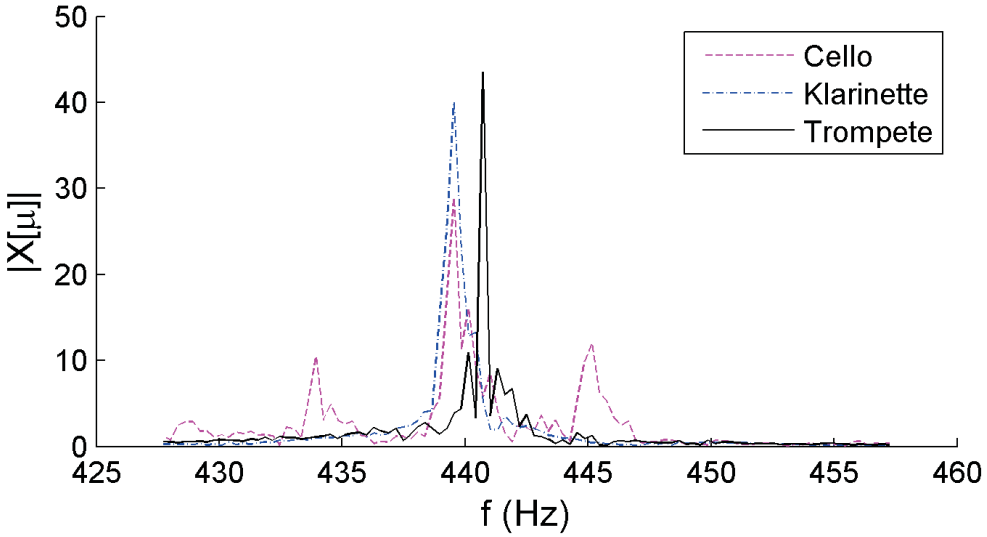

Abbildung 2.11: DFT-Amplitude des Kanals 15 bei realen Tönen (McGill University Master Samples (Eerola und Ferrer, 2008)) mit einer Grundfrequenz von 440 Hz dreier Musikinstrumente.

zwei Modelle getestet werden. In den Hauptexperimenten dieser Arbeit (Kapitel 6 und 7) werden daher die beiden Klassifikationsverfahren lineare *Support-Vektor-Maschine* (SVM) und *Random Forest* (RF) verwendet, die in Vorstudien auch mit Standardeinstellungen stabile Ergebnisse geliefert haben.

2.3.1 Entscheidungsbäume und Random Forest

Entscheidungsbäume gehören zu den intuitivsten Klassifikationsmodellen. Ihre Grundidee ist ein Modell, das aus einer Menge von hierarchischen Entscheidungsregeln besteht, die sich üblicherweise auf ein einziges Merkmal beziehen. Meistens sind diese Entscheidungen binär, so dass man das Modell durch einen binären Baum darstellen kann.

Wenn eine neue Beobachtung klassifiziert werden soll, geht man den Baum von oben herab und nimmt an jedem Knoten entweder die linke oder rechte Abzweigung in Abhängigkeit von der Entscheidungsregel des Knotens und dem entsprechenden Merkmalswert. Sobald ein Endknoten (Blatt) erreicht ist, wird ein Klassenlabel vergeben. Siehe Breiman (1996) für eine umfassendere Beschreibung von Entscheidungsbäumen.

Eine Erweiterung von Entscheidungsbäumen ist der *Random Forest*. Dieser kombiniert mehrere Entscheidungsbäume (siehe z.B. Breiman, 2001) Der Name *Random Forest*

kommt daher, dass die Konstruktion der Bäume zum gewissen Grad zufällig geschieht, beispielsweise wird jeder Baum nur mit einer zufälligen Teilmenge der Merkmale erzeugt.

2.3.2 Support-Vektor-Maschinen

Support Vektor Maschinen (SVMs) (Vapnik, 1998) gehören zu den derzeit besten Methoden für lineare und auch nichtlineare Klassifikation. Die lineare SVM trennt zwei Klassen, die mit einem Label $\psi \in \{-1, +1\}$ bezeichnet werden. Das Modell separiert die Klassen durch eine affin lineare Funktion $f(\vec{\phi}) = \vec{w}^T\vec{\phi} + b$, die durch einen gewichteten Vektor $\vec{w} \in \mathbb{R}^p$ und einem Bias oder Bestrafungsterm $b \in \mathbb{R}$ definiert ist. Eine neue Beobachtung $\vec{\phi}$ wird dann einfach durch das Vorzeichen bezüglich der Funktion f klassifiziert: $\text{sign}(f(\vec{\phi}))$.

Bei der Generierung eines SVM-Modells wird die affine lineare Funktion f gesucht, die einen Sicherheitsabstand zwischen den beiden Klassen maximiert. Da in vielen praktischen Problemen Ausreißer vorliegen, werden sogenannte Schlupfmerkmale ξ_i verwendet – für jeden Trainingspunkt eins –, wodurch die Stärke der jeweiligen Bedingungsverletzung gemessen wird:

$$\min_{\vec{w},b} \quad \frac{1}{2}\|\vec{w}\|^2 + C \cdot \sum_{i=1}^{n} \xi_i, \qquad \text{so dass } \psi_i \cdot (\vec{w}^T\vec{\phi}_i + b) \geqslant 1 - \xi_i \text{ und } \xi_i \geqslant 0. \tag{2.9}$$

Die Lösung $(\vec{w}^*, b^*)$ dieses Problems definiert das Standardmodell der linearen SVM. Sie besteht aus einem einzelnen (zu optimierenden) Parameter, $C > 0$, der den Trade-off zwischen der Maximierung der Abstände und der Minimierung der Abstandsverletzungen definiert.

Viele praktische Probleme, wie auch die hier untersuchten Musikklassifikationsprobleme, benötigen mehr als zwei Klassen ($G > 2$). Eine Möglichkeit die SVM dahingehend zu erweitern ist der *one-versus-one* Ansatz, der prinzipiell auch für jedes andere binäre Klassifikationsverfahren anwendbar ist. Dabei wird das G-Klassen Problem in $G(G-1)/2$ binäre Probleme konvertiert, wobei für jedes mögliche Paar von Klassen eine SVM darauf trainiert wird, genau diese beiden Klassen zu trennen. Eine neue Beobachtung wird dann von jeder SVM bezüglich ihrer beiden Klassen eingestuft. Zum Schluss wird die Beobachtung dann derjenigen Klasse zugeordnet die am häufigsten gewählt worden ist.

2.4 Merkmalsselektion

Die Merkmalsselektion wählt die wichtigsten Merkmale aus, wodurch für die Klassifikationsaufgaben, die mit der Ohrmodellausgabe arbeiten, erheblich Rechenzeit für den Ohrmodellprozess selbst, für die Merkmalsextraktion und für das Training des Klassifikationsmodells eingespart werden kann. Ein weiterer Vorteil der Merkmalsselektion ist, dass ein Klassifikationsmodell, das mit weniger Merkmalen generiert worden ist, in der Regel leichter zu interpretieren ist. Zu wissen, welche Merkmale wichtig sind, hilft auch dabei neue verbesserte Merkmale zu generieren. Weiterhin kann Merkmalsselektion sogar oft die Klassifikationsergebnisse verbessern, vor allem bei Klassifikationsmethoden, die Probleme mit redundanten und bedeutungslosen Merkmalen haben.

Das Problem der Merkmalsselektion ist jedoch ihre hohe Komplexität, so dass es in der Regel nicht möglich ist alle möglichen Variablenkombinationen auszuprobieren (was zudem auch das Risiko einer Überanpassung erhöht). Um die Rechenkomplexität zu reduzieren, gibt es zwei einfache Verfahren: Vorwärtsselektion und Rückwärtsselektion (Kohavi und John, 1997). Vorwärtsselektion ist ein „gieriges“ Suchverfahren, das mit einer leeren Merkmalsmenge startet und in jeder folgenden Iteration das Merkmal in diese Menge aufnimmt, das die Fehlerrate am stärksten vermindert. Es wird so lange iteriert bis die Hinzunahme keines Merkmals eine Verbesserung bringt die einen vorher festgelegten Schwellenwert übersteigt. Rückwärtsselektion funktioniert genau anders herum. Gestartet wird mit der vollständigen Merkmalsmenge und in jeder Iteration wird das Merkmal entfernt, das am wenigsten Zugewinn bezüglich der Fehlerrate bringt. Dabei wird üblicherweise auch eine leichte Verschlechterung zugelassen, als Kompensation dafür, dass das Modell einfacher wird.

Beide Verfahren haben eine Komplexität von $O(n^2)$, was bei den Verfahren mit den $\approx$ 1000 Ohrmodellmerkmalen immer noch eine zu hohe Rechenzeit bedeutet. Daher wird im Folgenden eine gruppenbasierte Variablenselektion eingeführt, die ausnutzt, dass die Merkmale durch zwei Dimensionen – (1) Merkmalstyp und (2) Kanalnummer – beschrieben werden können. Bei dieser Merkmalsselektion wird dann jede Merkmalsgruppe wie ein einziges Merkmal aufgefasst, wodurch die Rechenzeit deutlich reduziert wird. Die erste Möglichkeit ist es also alle Merkmale des gleichen Typs zusammenzufassen. Hierbei wird angenommen, dass der Typ entscheidend für die Wichtigkeit ist, unabhängig davon aus welchem Kanal das Merkmal kommt. Bei diesem Verfahren gib es 29 Gruppen für die Tonhöhenklassifikation (Vergleiche Tabelle 4.1 in Kapitel 4) und 21 Gruppen für die Instrumentenklassifikation (siehe Tabelle 5.1 in Kapitel 5). Bei der zweiten Möglichkeit

werden alle Merkmale, die im gleichen Kanal extrahiert worden sind, zusammengefasst. Hierbei wird angenommen, dass die Kanalnummer die entscheidendere Information ist und die Wichtigkeit eines Merkmals im Wesentlichen davon abhängig ist. Da das verwendete Ohrmodell (ohne simulierte Hörschädigung) aus 41 Kanälen besteht, ergeben sich für beide Musikklassifikationsanwendungen 41 Gruppen. Dieses zweite Verfahren hat den zusätzlichen Vorteil, dass hierdurch komplette Kanäle wegfallen, so dass deren Ausgabe bei der Ohrmodellsimulation erst gar nicht berechnet werden müssen und somit zusätzliche Rechenzeit eingespart wird. Für die Tonhöhenerkennung ist das zwar nur die halbe Wahrheit, da die Amplituden möglicher Teiltöne ($P_{pl}^{mean}[k]$ und $P_{pl}^{max}[k]$) über alle Kanäle berechnet werden, aber möglicherweise ist es ausreichend diese Merkmale über die verbliebenen Kanäle hinweg zu berechnen oder sie stellen sich sowieso als weniger wichtig heraus. In den Experimenten dieser Arbeit ist die mindeste Verbesserung pro Iteration für die Vorwärtsselektion auf 0.01 gesetzt und für die Rückwärtsselektion auf -0.001.

2.5 Modellbasierte Optimierung (MBO)

Sequentielle modellbasierte Optimierung (MBO) – auch als Efficient Global Optimization (EGO) (Jones, Schonlau und Welch, 1998) bezeichnet – ist eine immer populärer werdende Methode um parametrisierbare Optimierungsprobleme zu lösen. Sie ist dann sinnvoll, wenn die Auswertung einer einzelnen Parameterkonfiguration sehr teuer ist, so dass sich ein höherer Aufwand für die Vorhersage von vielversprechenden Konfigurationen lohnt. Ein typisches Beispiel eines solchen Problems ist die Parameterkonfiguration eines zeitaufwändigen Algorithmus. In dieser Arbeit wird MBO an zwei Stellen verwendet. Zum einen werden die Parameter des Algorithmus für die Einsatzzeiterkennung optimiert (Kapitel 3.2), und zum anderen wird MBO für die Optimierung eines Hörgerätealgorithmus eingesetzt (Kapitel 8). Des Weiteren gibt es aber auch noch andere Bereiche dieser Arbeit, für die der Einsatz von MBO in weiterführenden Forschungsarbeiten sinnvoll wäre. Beispielsweise könnte MBO dafür verwendet werden, verbesserte Merkmale für die Musikklassifikationsaufgaben zu generieren, die bislang üblicherweise relativ unsystematisch definiert sind. Ein weiterer möglicher Einsatzbereich für MBO ist die Hyperparameteroptimierung der Klassifikationsmodelle, wie dies in Lang (2015) bereits für die Hyperparameteroptimierung von Überlebenszeitmodellen durchgeführt worden ist.

Das Ziel der Optimierung ist die effiziente Suche nach einer möglichst guten Parameterkonfiguration $\boldsymbol{\theta}$ in dem Raum der zugelassenen Einstellungen Θ eines Algorithmus $\mathcal{A}$ bezüglich einer Kostenfunktion $c_{\mathcal{A}}$, die üblicherweise als Minimierungsproblem definiert ist (Lang, 2015). Für die Algorithmen dieser Arbeit ist die Kostenfunktion beispielsweise das negierte F-Maß (siehe Kapitel 3) für die Optimierung der Einsatzzeiterkennung oder die Fehlklassifikationsrate für die Hörgeräteoptimierung. Beide Optimierungsprobleme haben gemeinsam, dass der Algorithmus bezüglich einer großen Menge von Instanzen Ω ($\widehat{=}$ Musikstücke) optimiert werden soll. In dieser Arbeit wird davon ausgegangen, dass jede Instanz gleich wichtig ist, womit die Kostenfunktion als das arithmetische Mittel der Kosten aller Instanzen definiert wird. Sei $c_{\mathcal{A}}^{\omega} : \Omega \times \Theta \to \mathbb{R}$ die Einzelkostenfunktion bezüglich einer Instanz $\omega \in \Omega$. Dann gilt

$$c_{\mathcal{A}}(\Omega, \boldsymbol{\theta}) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} c_{\mathcal{A}}^{\omega}(\omega, \boldsymbol{\theta}). \tag{2.10}$$

Da für die Musikprobleme in dieser Arbeit die Anzahl der existierenden Instanzen viel zu hoch ist, kann der Algorithmus nur bezüglich einer Teilmenge $\Omega_{\text{opt}} \subset \Omega$ optimiert werden, die möglichst repräsentativ ausgewählt werden sollte. Damit lässt sich das mit MBO zu lösende Optimierungsproblem als

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}}\, c_{\mathcal{A}}(\Omega_{\text{opt}}, \boldsymbol{\theta}) \tag{2.11}$$

formulieren. Hierbei besteht die Gefahr, das eine Überanpassung an die Optimierungsmenge Ω_{opt} stattfindet. Die tatsächlichen Kosten $c_{\mathcal{A}}(\Omega, \boldsymbol{\theta})$ werden daher besser durch $c_{\mathcal{A}}(\Omega_{\text{eval}}, \boldsymbol{\theta})$ unter Verwendung einer unabhängigen Evaluierungsmenge $\Omega_{\text{eval}} \subset \Omega$ mit $\Omega_{\text{opt}} \cap \Omega_{\text{eval}} = \varnothing$ geschätzt. Analog zur Evaluierung von Klassifikationsverfahren kann die Optimierungsmenge Ω_{opt} auch als Trainingsmenge und die Evaluierungsmenge Ω_{eval} als Testmenge bezeichnet werden. Darauf wird in dieser Arbeit jedoch verzichtet um Verwechslungen zu vermeiden, da innerhalb der Optimierung des Hörgerätealgorithmus eine Klassifikation durchgeführt wird, die intern die Optimierungsmenge Ω_{opt} in Trainings- und Testmengen aufteilt (Kreuzvalidierung).

Zur Lösung von 2.11 wendet MBO eine iterative Vorgehensweise an. Das Ziel dabei ist, in jeder Iteration eine möglichst aussichtsreiche Parameterkonfiguration $\boldsymbol{\theta}$ – im Folgenden auch als Punkt bezeichnet – vorzuschlagen, unter Berücksichtigung aller bislang bekannten Funktionsauswertungen. Der grundlegende Aufbau von MBO ist in Abbildung 2.12 schematisch dargestellt. Nachdem zunächst für n_{init} initiale Punkte Θ_{init}

die tatsächlichen Kosten $c_{\mathcal{A}}(\Omega_{\text{opt}}, \boldsymbol{\theta})$ mit $\boldsymbol{\theta} \in \Theta_{\text{init}}$ ermittelt sind (Schritt 1), startet der iterative Prozess (Schritte 2 bis 4). Zunächst wird basierend auf den bisherigen n Punktauswertungen ein Regressionsmodell – auch als Surrogatmodell bezeichnet – erstellt, das die abhängige Zielvariable $y := c_{\mathcal{A}}(\Omega, \boldsymbol{\theta})$ durch die Parameter $\boldsymbol{\theta}$ erklärt. Mit diesem Modell kann jedem Punkt $\boldsymbol{\theta}$ eine erwartete Güte $\hat{c}_n(\boldsymbol{\theta})$ und deren geschätzte Unsicherheit $\hat{s}_n^2(\boldsymbol{\theta})$ zugeordnet werden (Schritt 2). Mit Hilfe dieser Maße quantifiziert ein sogenanntes Infill-Kriterium $\text{ei}_n(\boldsymbol{\theta})$ die Güte jedes Punktes $\boldsymbol{\theta}$ und ein Infill-Optimierer sucht einen dementsprechend möglichst aussichtsreichen Punkt $\boldsymbol{\theta}_{n+1} = \text{argmax}_{\boldsymbol{\theta} \in \Theta} \text{ei}_n(\boldsymbol{\theta})$ (Schritt 3). Für diesen Punkt werden daraufhin die tatsächlichen Kosten $c_{\mathcal{A}}(\Omega_{\text{opt}}, \boldsymbol{\theta}_{n+1})$ berechnet (Schritt 4). Anschließend wird die nächste Iteration $n + 1$ gestartet, in der dann zunächst das Surrogatmodell $\hat{c}_{n+1}(\cdot)$ unter Berücksichtigung von $c_{\mathcal{A}}(\Omega_{\text{opt}}, \boldsymbol{\theta}_{n+1})$ aktualisiert wird. Dieser iterative Prozess endet, sobald ein vorher definiertes Abbruchkriterium, z.B. ein maximales Budget, erreicht ist.

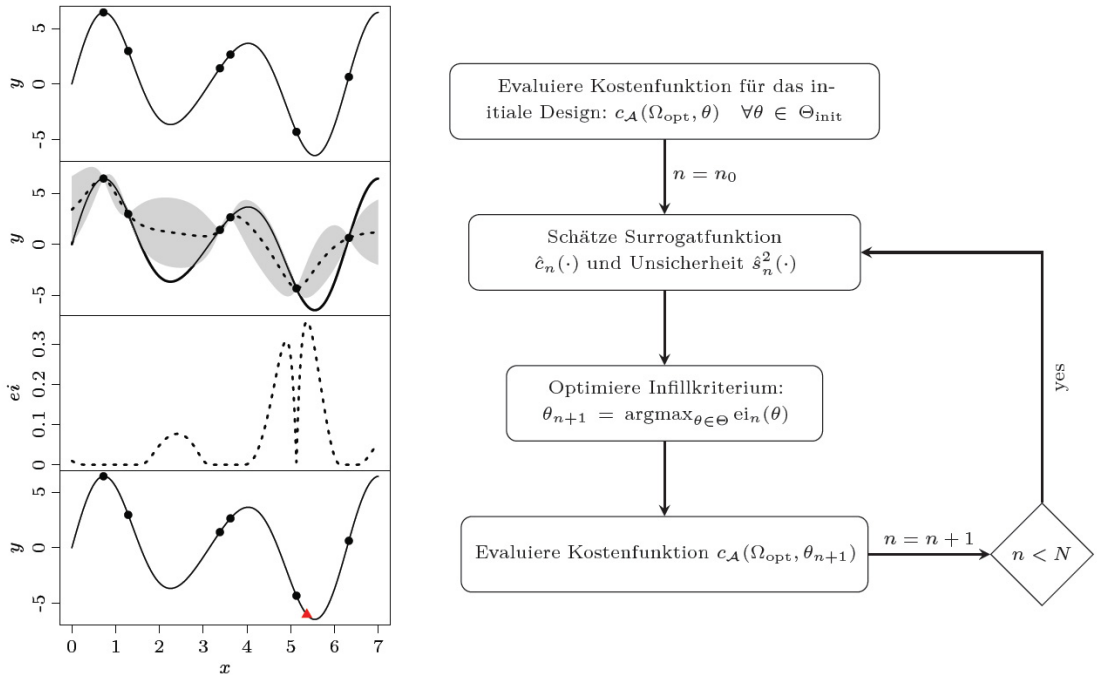

Abbildung 2.12: Funktionsweise der modellbasierten Optimierung (Weihs u. a., 2016).

Wie anhand von Abbildung 2.12 zu sehen ist, gibt es für die MBO-Schritte selbst verschiedene Variationen und Einstellmöglichkeiten, die im folgenden kurz erläutert werden. Die initial auszuwertenden n_{init} Punkte werden mittels eines initialen Designs

bestimmt, wobei n_{init} möglichst minimal sein sollte, aber doch groß genug, um ein sinnvolles Surrogatmodell schätzen zu können. Empfohlen wird $n_{\text{init}} \in \{5p, 6p, \ldots, 10p\}$, wobei p die Anzahl der zu optimierenden Parameter ist (Weihs u. a., 2016). Prinzipiell kann das Design völlig zufällig erstellt werden, üblicherweise wird jedoch ein *Latin-Hypercube-Sampling* (LHS, Stein (1987)) bevorzugt, um den Suchraum möglichst gleichmäßig abzudecken. Als Surrogatmodell sind Kriging-Modelle (Krige, 1951) am verbreitetsten, die den Vorteil haben, dass sie neben der Modellvorhersage $\hat{c}_n(\boldsymbol{\theta})$ auch direkt die Unsicherheit $\hat{s}_n^2(\boldsymbol{\theta})$ mitliefern. Das populärste Infill-Kriterium ist das *Expected Improvement* (erwartete Verbesserung, EI). Die Verbesserung I eines Punktes $\boldsymbol{\theta}$ ist durch

$$I(\boldsymbol{\theta}) = \max\left(c_{\mathcal{A}}^{\min} - c_{\mathcal{A}}(\Omega_{\text{opt}}, \boldsymbol{\theta}), 0\right) \tag{2.12}$$

gegeben, wobei $c_{\mathcal{A}}^{\min}$ der beste bisher ausgewertete Funktionswert ist. Demnach entspricht $I(\boldsymbol{\theta})$ der positiven Kostenreduzierung des Punktes $\boldsymbol{\theta}$, mit einer Deckelung bei 0, falls er schlechter als $c_{\mathcal{A}}^{\min}$ ist. Nach Jones, Schonlau und Welch (1998) ist dann für ein Kriging-Modell die erwartete Verbesserung $E[I(\boldsymbol{\theta})]$ eines Punktes $\boldsymbol{\theta}$ durch

$$E\left[I(\boldsymbol{\theta})\right] = \left(c_{\mathcal{A}}^{\min} - \hat{c}_n(\boldsymbol{\theta})\right) \Phi\left(\frac{c_{\mathcal{A}}^{\min} - \hat{c}_n(\boldsymbol{\theta})}{\hat{s}_n(\boldsymbol{\theta})}\right) + \hat{s}_n(\boldsymbol{\theta}) \phi\left(\frac{c_{\mathcal{A}}^{\min} - \hat{c}_n(\boldsymbol{\theta})}{\hat{s}_n(\boldsymbol{\theta})}\right) \tag{2.13}$$

gegeben, wobei $\Phi(\cdot)$ die Verteilungsfunktion und $\phi(\cdot)$ die Dichte der Standardnormalverteilung sind. Das *Expected Improvement* stellt somit einen Kompromiss zwischen der Modellvorhersage $\hat{c}_n(\boldsymbol{\theta})$ (Exploitation) und der Unsicherheit $\hat{s}_n^2(\boldsymbol{\theta})$ (Exploration) dar.

Der optimale Punkt $\boldsymbol{\theta}^*$ bezüglich der erwarteten Verbesserung kann nicht direkt berechnet werden, sondern muss stattdessen durch eine Infill-Optimierung ermittelt werden. Die Auswertung eines Punktes beruht an dieser Stelle aber nur noch auf der Funktionsauswertung des Surrogatmodells, was nur eine sehr geringe Rechenzeit erfordert. Daher kommen beispielsweise Evolutionäre Algorithmen für die Optimierung in Frage, in dieser Arbeit wird jedoch die durch das R-Paket mlrMBO propagierte *Focus Search* verwendet (Lang, 2015; Weihs u. a., 2016).

Üblicherweise durchläuft MBO eine vorher festgelegte Anzahl von n_{iter} Iterationen. Zum einen ist so der Endzeitpunkt der Optimierung abschätzbar, und häufig steht nur ein maximales Zeitbudget zur Verfügung. Ein weiterer Vorteil ist aber auch die bessere Vergleichbarkeit zwischen den Ergebnissen verschiedener Optimierungsläufe. Für die Anwendung selbst hat diese strikte Begrenzung jedoch Nachteile, denn das Konvergenzverhalten der Optimierung wird nicht berücksichtigt. Um dies zu verhindern,

kann ein qualitatives Abbruchkriterium verwendet werden. Eine Möglichkeit besteht z.B. darin, die Optimierung zu beenden, sobald das Infill-Kriterium einen definierten (relativen) Schwellenwert unterschreitet (Huang u. a., 2006).

Am Ende der Optimierung liefert MBO die ausgewerteten Punkte $\boldsymbol{\theta}_n$ für $n \in \{1, \ldots, n_{\text{iter}}\}$ mit ihren dazugehörigen Funktionswerten $c_{\mathcal{A}}(\Omega_{\text{opt}}, \boldsymbol{\theta}_n)$. Aus Anwendungssicht gilt in erster Linie das Interesse dem besten gefundenen Punkt $\boldsymbol{\theta}^*_{\text{MBO}}$, der durch

$$\boldsymbol{\theta}^*_{\text{MBO}} = \underset{n \in \{1,\ldots,n_{\text{iter}}\}}{\operatorname{argmin}} \; c_{\mathcal{A}}(\Omega_{\text{opt}}, \boldsymbol{\theta}_n) \tag{2.14}$$

definiert ist. Für eine verlässliche Schätzung der tatsächlichen Kosten bezüglich aller Instanzen, ist anschließend noch eine Auswertung auf einer unabhängigen Validierungsmenge Ω_{eval} zu empfehlen. Die tatsächlichen Kosten des optimierten Algorithmus werden dann durch $c_{\mathcal{A}}(\Omega_{\text{eval}}, \boldsymbol{\theta}^*_{\text{MBO}})$ geschätzt. Um analysieren zu können, wie diese Kosten mit der Charakteristik der Instanzen (hier: Musikstücke) zusammenhängt, ist es auch sinnvoll die Einzelkosten der Instanzen $\omega \in \Omega_{\text{eval}}$ zu betrachten. Schließlich ist aus Entwicklungssicht der genaue Optimierungsverlauf interessant, um die Hyperparametereinstellungen von MBO für spätere Läufe zu verbessern. Auf Grund der probabilistischen Eigenschaften von MBO sollten im Idealfall alle Optimierungsläufe mehrmals gestartet werden.

3 Einsatzzeiterkennung

Aufgabe der Einsatzzeiterkennung ist es, in einem Musikstück alle Zeitpunkte zu identifizieren, in denen ein neuer Ton beginnt. Dagegen sollen in der predominanten Einsatzzeiterkennung nur die Zeitpunkte identifiziert werden, in denen ein neuer Ton der Melodiestimme beginnt. Dieser Spezialfall wurde bislang in der Literatur noch nicht untersucht. Daher werden in dieser Arbeit, für eine bessere Vergleichbarkeit, auch die Einsatzzeiterkennung für monophone Musik, sowie die übliche polyphone Variante, die nach allen Toneinsätzen sucht, betrachtet. Für die predominante Variante wird der herkömmliche Algorithmus verwendet, dessen Parameterwerte jedoch durch eine Optimierung mit anderen Zielwerten (nur die Toneinsätze der Melodie) modifiziert werden.

Eine Schwierigkeit der Einsatzzeiterkennung besteht in der Unterscheidung zwischen wahren Tonanfängen und allmählichen Veränderungen und Modulationen innerhalb eines Tons, die je nach Musikinstrument, Spielweise und Raumumgebung sehr unterschiedlich sind. Man kann zwischen zwei Varianten der Einsatzzeiterkennung unterscheiden, je nachdem ob Entscheidungen direkt in Realzeit erfolgen müssen (online) oder ob auch die zukünftige Entwicklung des Signals mitberücksichtigt werden darf (offline). Für ein Verfahren, das als Ersatz für die menschliche Wahrnehmung gedacht ist, erscheint erst einmal die Online-Variante natürlicher, jedoch berücksichtigt auch unser Gehirn eine kurze Zeitspanne der Zukunft für die Hörwahrnehmung.[1] Dass die Wahrnehmung des Menschen nicht vollkommen chronologisch verläuft, zeigt beispielsweise das Phänomen der Rückwärtsmaskierung. Bei dieser wird ein Ton durch einen zweiten Ton, der später einsetzt (bis etwa 0.1 s), aber auf Grund einer höheren Lautstärke schneller verarbeitet wird, vollständig unterdrückt (Raab, 1963). In dieser Arbeit wird die Offline-Variante untersucht. Diese Entscheidung wurde auch unter Berücksichtigung der Aspekte getroffen, dass die automatische Einsatzzeiterkennung generell noch nicht zufriedenstellend funktioniert (im Vergleich zur menschlichen Wahrnehmungsfähigkeit) und die Online-Bedingung die Erkennung noch deutlich erschwert.

[1] Dies ist natürlich nur auf Grund einer kurzen Verzögerung (Latenz) der Wahrnehmung möglich.

3.1 Klassischer Algorithmus der Einsatzzeiterkennung

Die Mehrzahl der Algorithmen zur Einsatzzeiterkennung bestehen aus einer optionalen Vorverarbeitung, gefolgt von einer Reduktionsfunktion, durch welche die Anzahl der Abtastzeitpunkte reduziert wird, und einem Algorithmus zur Lokalisierung der Toneinsatzpunkte (englisch: *Peak-Picking*) (Bello u. a., 2005). In Bauer u. a. (2016) haben wir eine Vielzahl von möglichen Verfahren zu einem einzelnen Algorithmus zusammengefasst, der acht frei einstellbare Parameter beinhaltet. Anschließend haben wir das global beste Verfahren mit Hilfe der sequentiellen modellbasierten Optimierung (MBO) ermittelt. Dieser Algorithmus wurde von Bauer (2016) noch durch weitere Paramater ergänzt, und diese Erweiterung dient auch in dieser Arbeit als Grundlage. Allerdings wird er leicht vereinfacht, indem kategorielle Parameter auf konstante Werte festgesetzt werden, da diese standardmäßig von MBO nicht unterstützt werden. Es existieren zwar Erweiterungsmöglichkeiten, mit denen auch kategorielle Parameter optimiert werden können (Bauer, 2016), doch ist nicht abschließend geklärt, ob diese auch unter allen Bedingungen akzeptabel funktionieren. Da sich in den oben aufgeführten Arbeiten jedoch auch herausgestellt hat, dass die kategoriellen Parameter entweder keinen großen Einfluss auf die Güte des Algorithmus haben oder aber ein bestimmter Wert klar dominierend ist, wurde hier, nicht zuletzt auch aus Rechenzeitgründen, auf die Optimierung dieser Parameter verzichtet. Die festgelegten Konstanten sind aus den Ergebnissen in Bauer (2016) motiviert und werden in der folgenden Beschreibung des Algorithmus im Einzelnen begründet.

Der klassische Algorithmus der Einsatzzeiterkennung besteht aus 7 Schritten, die in Abbildung 3.1 veranschaulicht sind. Die zu jedem Schritt dazugehörigen freien Parameter – die später optimiert werden sollen – sind in Klammern angegeben.

Im ersten Schritt wird das Eingangssignal in kleine Fenster unterteilt, deren Fenstergröße M (Anzahl an Abtastwerten) ist. Diese Fenster überlappen sich gewöhnlich, wobei der Abstand zwischen zwei benachbarten Fenstern durch die Sprungweite h definiert wird. Für jedes Fenster wird mit Hilfe der diskreten Fourier Transformation (DFT, siehe Gleichung 2.7 in Kapitel 2.2) die spektrale Amplitude $|X[\mu]|$ berechnet.

Eine Fensterfunktion wird benötigt, um die Bereiche an den Rändern abzuschwächen, da die dortigen Werte sonst zu Artefakten führen können. In Bauer (2016) werden vier Fensterfunktionen untersucht: Rechteck-, Hanning-, Blackman- und Gauß- Fensterfunktion. Am verbreitetsten ist dabei das Hanning-Fenster, das auch in den dortigen Experimenten

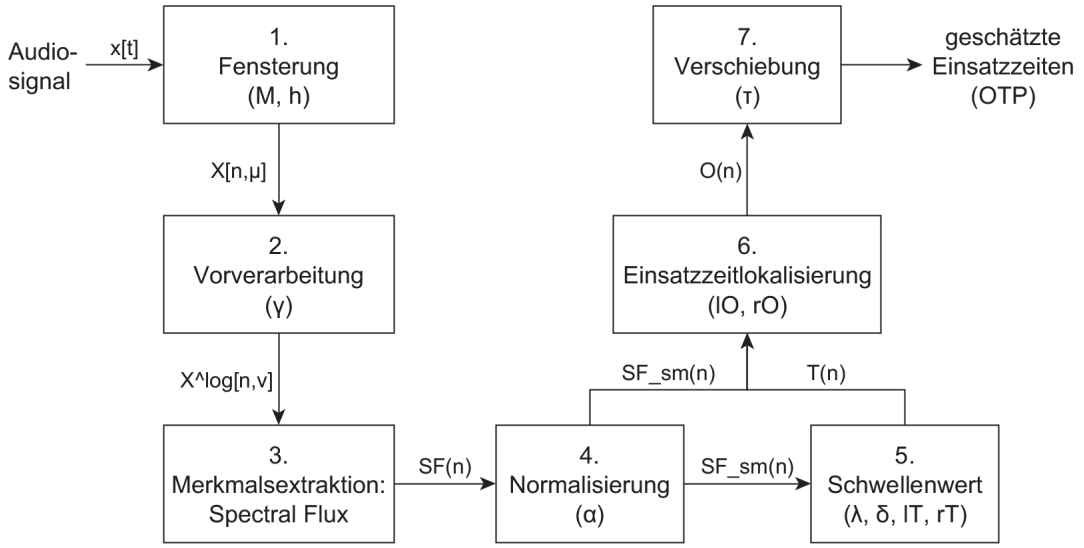

Abbildung 3.1: Blockdiagramm des klassischen Algorithmus der Einsatzzeiterkennung (ohne Ohrmodell).

am besten abschneidet. Daher wird in dieser Arbeit ausschließlich diese Fensterfunktion verwendet, wodurch der erste kategorielle Parameter vermieden wird. Sie ist definiert durch

$$w(t) = 0.5\left(1 - \cos\left(\frac{2\pi(t-1)}{M-1}\right)\right), \quad t = 1, \ldots, M, \tag{3.1}$$

wobei t den (Abtast-)Zeitpunkt bezeichnet. Da hier die DFT für jedes Fenster einzeln berechnet wird (Short Time Fourier Transform, STFT), ergibt sich somit insgesamt die Definition:

$$X_{\text{stft}}[n, \mu] = \sum_{k=0}^{M-1} x[h(n) + k]w[k] \exp\left(-i\frac{2\pi\mu k}{M}\right), \quad \mu = 0, \ldots, M-1\,, \tag{3.2}$$

wobei n den Fensterindex und μ den Frequenzindex bezeichnet. Die Funktion $h(n)$ definiert den ersten Abtastzeitpunkt des Fensters n. Im Folgenden ist nur noch die spektrale Amplitude $|X_{\text{stft}}[n, \mu]|$ der STFT von Interesse.

Im nächsten Schritt werden zwei Vorverarbeitungsschritte durchgeführt. Zunächst wird das Spektrum mit Hilfe einer Filterbank $F[\cdot]$ gefiltert, die in Böck, Krebs und Schedl (2012) eingeführt wurde. Diese bündelt die spektrale Amplitude gemäß der Tonhöhenskala

der westlichen Musik. Das gefilterte Spektrum ist definiert durch

$$X_{\text{filt}}[n,\nu] = \sum_{\mu=1}^{M} |X_{\text{stft}}[n,\mu]| \cdot F[\mu,\nu], \tag{3.3}$$

wobei ν den Index dieser Skala bezeichnet, die aus insgesamt $B = 82$ Frequenzlinien besteht. Der zweite Vorverarbeitungsschritt ist eine Logarithmierung dieses Spektrums, gemäß Eyben u. a. (2010):

$$X^{\log}[n,\nu] = \log(\gamma \cdot X_{\text{filt}}[n,\nu] + 1), \tag{3.4}$$

wobei γ ein zu optimierender Kompressionsparameter ist. In Bauer (2016) werden auch noch zwei boolesche Parameter eingeführt, durch die entschieden wird, ob die Vorverarbeitungsschritte 3.3 bzw. 3.4 überhaupt durchgeführt werden. Da die dortigen Ergebnisse darauf hindeuten, dass beide Schritte vorteilhaft sind, wird in dieser Arbeit jedoch auf diese Parameter verzichtet, und beide Schritte werden immer ausgeführt.

Anschließend wird für jedes Fenster ein Merkmal berechnet. Die Merkmale aller Fenster können dann zu einer vereinfachten Zeitreihe zusammengefasst werden, deren Länge im Vergleich zum ursprünglichen Signal $x[t]$ um den Faktor h (Sprungweite) reduziert ist. Die grundlegende Idee ist es, die Zeitpunkte dieser Reihe zu identifizieren, die einen Schwellenwert überschreiten und die außerdem ein lokales Maximum sind. Verschiedenster solcher Merkmale wurden vorgeschlagen, in Bauer (2016) werden insgesamt 18 verschiedene Merkmale getestet. In dieser Arbeit wird jedoch nur das gemäß aller aktuellen Vergleichsstudien (Böck, Krebs und Schedl, 2012; Rosao, Ribeiro und De Matos, 2012; Bauer u. a., 2016; Bauer, 2016) erfolgreichste Merkmal verwendet, wodurch ein weiterer kategorieller Parameter aus der Optimierung entfernt werden kann. Dieses Merkmal heißt *Spectral Flux* (SF) (deutsch: spektraler Fluss) und beschreibt den positiven spektralen Unterschied der jeweils benachbarten Fenster:

$$\begin{aligned} SF(n) = \sum_{\nu=1}^{B} H(X^{\log}[n,\nu] - X^{\log}[n-1,\nu]) \\ \text{mit } H(x) = (x + |x|)/2. \end{aligned} \tag{3.5}$$

Dabei stellt der Filter H sicher, dass nur Anstiege der spektralen Amplitude berücksichtigt werden, denn Abstiege deuten eher auf das Ende eines Tons hin.

Das Ziel der Normalisierung in Schritt 4 ist die Transformation des Merkmalsvektors in

eine standardisierte Form, damit das anschließende Schwellenwertverfahren besser für verschiedene Musikstücke mit unterschiedlichen Dynamiken funktionieren kann. Dafür wird eine exponentielle Glättung angewendet, die durch

$$\begin{aligned} &SF_{\text{sm}}(1) = SF(1) \quad \text{und} \\ &SF_{\text{sm}}(n) = \alpha \cdot SF(n) + (1-\alpha) \cdot SF_{\text{sm}}(n-1), \quad \text{für } n = 2, \ldots, L \end{aligned} \tag{3.6}$$

definiert ist, wobei L die Anzahl der Fenster bezeichnet, und $\alpha \in [0, 1]$. Für $\alpha = 1$ bleiben alle Werte der Zeitreihe unverändert, während für $\alpha = 0$ jedes Element den Wert von $SF(1)$ übernimmt.

Es hat sich herausgestellt, dass ein globaler Schwellwert für die Identifizierung der richtigen Zeitpunkte zu unflexibel ist, da sich die Dynamik in vielen Musikstücken öfters ändert. Daher ist ein lokaler Schwellenwert, der sich auch aus den Werten in der Umgebung ableitet, besser geeignet. Die verwendete Schwellenwertfunktion besteht aus einem globalen Bestandteil δ und einem lokalen Bestandteil, der durch den Parameter λ gewichtet wird. Sie ist definiert durch (Rosao, Ribeiro und De Matos, 2012)

$$T(n) = \delta + \frac{\lambda}{l_T + r_T + 1} \sum_{i=-l_T}^{r_T} SF_{\text{sm}}(n+i), \quad \text{für } n = 1, \ldots, L, \tag{3.7}$$

wobei l_T und r_T die Anzahl an Fenstern links bzw. rechts des aktuellen Fensters bezeichnen. Durch diese Parameter wird der lokale Bereich definiert, der für die Berechnung des Schwellenwerts zum Zeitpunkt n verwendet wird. Anstatt des Mittelwertes können auch andere Schwellenwertfunktionen verwendet werden. Beispielsweise werden in Bauer (2016) auch Median und Quantil getestet, jedoch schneidet dabei der Mittelwert am besten ab, der hier ausschließlich verwendet wird.

Zum Schluss werden die lokalisierten Toneinsatzzeiten gemäß der bereits erwähnten Bedingungen selektiert: Sie müssen den lokalen Schwellenwert überschreiten und ein lokales Maximum sein:

$$O(n) = \begin{cases} 1, & \text{falls } SF_{\text{sm}}(n) > T(n) \quad \text{und} \\ & SF_{\text{sm}}(n) = \max(SF_{\text{sm}}(n - l_O), \ldots, SF_{\text{sm}}(n + r_O)) \\ 0, & \text{sonst.} \end{cases} \tag{3.8}$$

$\boldsymbol{O} = (O(1), \ldots, O(L))^T$ ist der Toneinsatzvektor und l_O bzw. r_O sind Parameter, welche die Anzahl an Fenstern nach links bzw. rechts angeben, wodurch der Bereich für die

Maximumsbedingung definiert ist. Fenster mit $O(n) = 1$ werden in Zeitangaben umgerechnet, die alle zusammen die geschätzten Einsatzzeiten ergeben. Zum Schluss werden diese Zeiten noch einmal um eine kleine Zeitkonstante τ verschoben, um systematische Latenzaspekte des Algorithmus kompensieren zu können, falls dieser die Einsatzzeiten im Schnitt entweder zu spät oder zu früh findet.

Für die korrekte Erkennung eines Toneinsatzes ist eine kleine Toleranz erlaubt, die in dieser Arbeit auf ±25 ms festgelegt ist. Die Leistung der Einsatzzeiterkennung wird üblicherweise mit dem F-Maß gemessen (Dixon, 2006):

$$F = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \quad F \in [0, 1], \tag{3.9}$$

wobei TP die Anzahl der korrekt erkannten Toneinsätze, FP die Anzahl der zusätzlichen falsch prognostizierten Einsätze und FN die Anzahl der übersehenen Einsätze sind. $F = 1$ entspricht einer perfekten Erkennung, wohingegen $F = 0$ bedeutet, dass kein Einsatz korrekt erkannt worden ist. Abgesehen von diesen beiden Extremwerten, sind die Werte des F-Maßes jedoch schwierig zu interpretieren, denn es ist nicht direkt klar, wie hoch die einzelnen Fehlerraten sind. Daher wird im Folgenden die Anzahl an Fehlern beispielhaft für drei spezielle Szenarien in Abhängigkeit vom F-Wert und der Anzahl an wahren Toneinsätzen $C_{\text{true}} = TP + FN$ aufgelöst:

$$FP = 0 \quad \Longrightarrow \quad FN = \left(1 - \frac{F}{2 - F}\right) \cdot C_{\text{true}} \tag{3.10}$$

$$FN = 0 \quad \Longrightarrow \quad FP = \left(\frac{2}{F} - 2\right) \cdot C_{\text{true}} \tag{3.11}$$

$$FP = FN \quad \Longrightarrow \quad FP = FN = (1 - F) \cdot C_{\text{true}} \tag{3.12}$$

Für einen F-Wert von 0.9 bedeutet dies im Fall von $FP = 0$, dass 18.2% aller Einsätze unentdeckt bleiben, oder falls $FN = 0$ gilt, dass $22.2\% \cdot C_{\text{true}}$ falsche Einsätze gefunden werden. Im Fall von $FP = FN$ werden 10% aller Einsätze nicht gefunden und noch einmal genauso viele falsche Zeitpunkte zu viel.

3.2 Parameteroptimierung der Einsatzzeiterkennung

Das im vorigen Abschnitt beschriebene Verfahren zur Einsatzzeiterkennung enthält insgesamt 11 zu optimierende Parameter, die in Tabelle 3.1 aufgelistet sind. Die für die

Parametername	Minimaler Wert	Maximaler Wert
Fenstergröße M	2^{10}	2^{12}
Sprungweite h	400	1600
γ	0.01	20
α	0	1
λ	0	2.6
δ	0	10
l_T	0 s	0.5 s
r_T	0 s	0.5 s
l_O	0 s	0.25 s
r_O	0 s	0.25 s
Verschiebung τ	-0.025 s	0.025 s

Tabelle 3.1: Parameter der klassischen Einsatzzeiterkennung und ihre untersuchten Wertebereiche.

Experimente in Kapitel 6 und 7 vorgegebenen Intervalle sind durch Vorversuche motiviert. Für die Fenstergröße werden aus Rechenzeitgründen nur Zweierpotenzen zugelassen (*Fast Fourier Transform*, FFT), obwohl für zukünftige Untersuchungen auch andere Werte interessant wären. Bei der verwendeten Abtastrate von 44100 liegt die Fenstergröße somit zwischen 23 ms (2^{10} Abtastwerte) und 93 ms (2^{12} Abtastwerte). Die Sprunggröße zwischen zwei Fenstern liegt zwischen 9 ms (400 Abtastwerte) und 36 ms (1600 Abtastwerte), wobei sie maximal die Fenstergröße erreichen darf und gegebenenfalls auf deren Wert reduziert wird. Die Definitionsbereiche der Parameter l_T, r_T, l_O und r_O sind für die Optimierung in Sekunden angegeben, werden aber für die oben aufgeführten Definitionen in die Anzahl von Fenstern umgerechnet, die abhängig von Fenstergröße und Sprungweite sind.

Zur Optimierung des Verfahrens wird eine Parametereinstellung gesucht, die, basierend auf einer Menge von Trainingsdaten, einen möglichst hohen mittleren F-Wert erreicht. Durch diese Optimierung kann das klassische Verfahren auch auf die predominante Variante und auch auf die Ohrmodellausgabe angepasst werden. Da die Auswertung einer Parameterkonfiguration sehr zeitaufwändig ist (mehrere Minuten auf dem verwendeten Linux-HPC-Cluster-System[2]), wird hierfür sequentielle modellbasierte Optimierung (MBO) verwendet (siehe Kapitel 2.5).

[2] `http://lidong.itmc.tu-dortmund.de/ldw/index.php?title=System_overview&oldid=259`

3.3 Einsatzzeiterkennung mit Hilfe des Ohrmodells

Statt auf der akustischen Wellenform $x[t]$, kann der in den vorigen Abschnitten beschriebene Algorithmus zur Einsatzzeiterkennung auch auf der Ausgabe eines Ohrmodellkanals $p[t, k]$ ausgeführt werden. Um den Algorithmus individuell für jeden Kanal anzupassen, werden die Parameter wieder mittels MBO optimiert. Das Ergebnis dieses Prozesses ist dann für jeden Kanal k eine geordnete Menge von geschätzten Einsatzzeitpunkten OTP_k. Hierbei entsteht nun die Problematik, wie diese verschiedenen Schätzungen zu einer Gesamtschätzung aggregiert werden können. Die einfachste Methode ist es, nur den Kanal zu betrachten, der in der Trainingsphase das beste Ergebnis erzielt. Jedoch ist diese Methode suboptimal, denn hierbei bleibt die Information aller anderen Kanäle unberücksichtigt.

Ein fast identisches Problem entsteht bei der Erweiterung des klassischen Algorithmus, bei der nicht mehr nur ein Merkmal (*Spectral Flux*), sondern mehrere Merkmale berücksichtigt werden sollen. Im Unterschied zu herkömmlichen Klassifikationsproblemen, bei denen man prinzipiell so viele Merkmale wie gewünscht verarbeiten kann, müssen bei der Einsatzzeiterkennung zeitliche Abhängigkeiten berücksichtigt werden. Da das F-Maß bezüglich eines Toleranzbereichs definiert wird, ist der wahre Wert eines geschätzten Toneinsatzes auch von den anderen geschätzten Zeitpunkten abhängig. Beispielsweise kann ein richtig geschätzter Zeitpunkt durch die Hinzunahme eines weiteren Zeitpunktes, der den wahren Toneinsatz noch etwas genauer schätzt, auf einmal falsch werden. Durch diese Problematik wird der Einsatz herkömmlicher Klassifikationsverfahren nicht unmöglich, aber dahingehend erschwert, dass zusätzlich ein Mechanismus benötigt wird, der benachbarte Toneinsatzschätzungen zusammenfasst. Prinzipiell gibt es zwei Vorgehensweisen:

1. Zusammenfassung der Merkmalsvektoren vor der Lokalisation der Toneinsätze,

2. Zusammenfassung der individuellen Schätzungen der Toneinsätze (nach der Lokalisation).

Der Vorteil der ersten Variante ist, dass dabei die Information verwendet wird, die durch die exakten Merkmalswerte gegeben ist. Dagegen ist die zweite Variante flexibler, denn in der ersten Variante müssen Fenstergrößen und Sprungweiten aller Einzelschätzungen identisch sein. Auch verschiedene Latenzzeiten der Schätzungen können in der zweiten Variante viel einfacher berücksichtigt werden, denn τ wird unabhängig für jede Schätzung gesetzt.

In Bauer u. a. (2014) haben wir ein Verfahren für die erste Vorgehensweise vorgeschlagen und getestet, welches die einzelnen Merkmalsvektoren per Quantil aggregiert. Der dort verwendete Algorithmus zur Einsatzzeiterkennung ist allerdings deutlich vereinfacht. Statt des hier verwendeten Merkmals *Spectral Flux*, wird dort eine Linearkombination verwendet, die mittels einer Gewichtung $w_A \in [0, 1]$ die Unterschiede der Amplitudenmaxima (F_A) und den Korrelationskoeffizienten der Spektren (F_S) zweier aufeinanderfolgender Fenster kombiniert:

$$CombF = w_A \cdot F_A + (1 - w_A) \cdot F_S. \tag{3.13}$$

Dieser Vektor wird für jeden Kanal unabhängig berechnet und anschließend wird zu jedem Zeitpunkt über alle Kanäle das p-Quantil ermittelt. Daraus ergibt sich ein Gesamtvektor, aus dem dann mittels eines globalen Schwellenwerts die Toneinsätze lokalisiert werden. Für p werden drei verschiedene Werte getestet ($p \in \{0.05, 0.5, 0.95\}$), wobei $p = 0.95$ am besten abschneidet. Die Ergebnisse zeigen, dass dieses Verfahren in vielen Fällen besser abschneidet als das Vergleichsverfahren ohne Ohrmodell, jedoch stark vom Musikinstrument abhängt. Für Klavier, Klarinette und Violine funktioniert das Verfahren sehr gut, dagegen schneidet es für Trompete und Flöte schlechter ab als das Vergleichsverfahren. Die Hauptschwäche des Verfahrens ist, dass sich Toneinsätze in den höheren Kanälen erst später bemerkbar machen, wie in Abbildung 3.2 beispielhaft für die Trompete gezeigt wird.

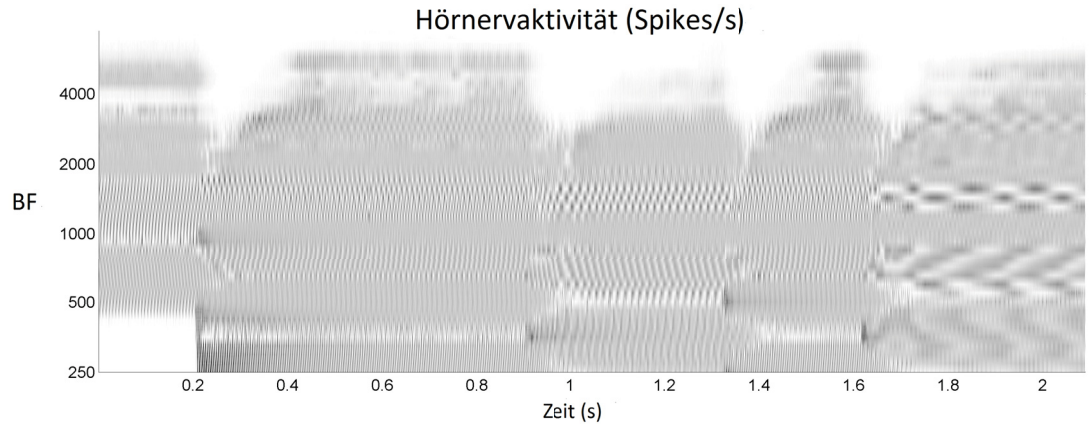

Abbildung 3.2: Ausschnitt der Ohrmodellausgabe eines Trompetenstücks. Bei den höheren Kanälen ist eine deutliche Reaktionsverzögerung zu erkennen (Bauer u. a., 2014).

Eine weitere Möglichkeit ist die multivariate Einsatzzeiterkennung, die in Bauer (2016) vorgestellt wird. Es wird ein Klassifikationsmodell gelernt, das für jedes Fenster ($\widehat{=}$ Beob-

achtung) schätzt, ob ein Toneinsatz vorliegt. Dafür werden 18 Merkmale (unter anderem *Spectral Flux*) verwendet, wobei neben den aktuellen Merkmalswerten auch Werte aus der näheren Vergangenheit und Zukunft (in der Offline-Variante) berücksichtigt werden. Mit diesem Klassifikationsmodell wird dann für jedes Fenster die Wahrscheinlichkeit eines Toneinsatzes vorhergesagt. Bei herkömmlichen Klassifikationsaufgaben würde nun einfach ein Schwellenwert (z.B. 0.5) verwendet, der jedem Fenster die Klasse „Toneinsatz“ oder „kein Toneinsatz“ zuweist. Hier muss jedoch auch noch die erwähnte zeitliche Abhängigkeit berücksichtigt werden, weshalb nur solche Fenster als „Toneinsatz“ klassifiziert werden, deren Wahrscheinlichkeitswert zudem ein lokales Maximum ist. Diese Regel ist somit fast identisch zu Gleichung 3.8, nur dass hier anstatt der Merkamalsausprägung die Wahrscheinlichkeitsschätzung des Klassifikationsmodells betrachtet wird. Zudem wird hier ein globaler Schwellenwert verwendet, da die Dynamik der Umgebung bereits im Klassifikationsmodell berücksichtigt ist. Unter Verwendung eines *Random Forests* für die Klassifikation erzielt das Verfahren für die Online-Variante eine Verbesserung des mittleren F-Wertes von 0.026 und in der Offline-Variante von 0.037 gegenüber der klassischen univariaten Variante mit *Spectral Flux*. Ein Problem des Verfahrens ist jedoch ein sehr hoher Rechenzeitbedarf, wofür vor allem die extrem große Anzahl von Beobachtungen verantwortlich ist, die jedoch auf Grund der sehr unausgeglichenen Klassenverteilung (fast alle Beobachtungen sind „kein Toneinsatz“) in der Trainingsphase benötigt wird. Für die Ohrmodellausgabe erweist sich das Verfahren daher als nicht praktikabel, denn hier steigt der Rechenzeitbedarf durch die höhere Anzahl von Merkmalen noch weiter an.

In Klapuri (1999) wird eine Methode für die zweite Aggregationsvariante vorgeschlagen, bei der die Einzelschätzungen erst am Ende kombiniert werden. In dem Artikel werden die Schätzungen von 21 nicht-überlappenden Bändern (Kanäle) einer Filterbank zusammengefasst. Zunächst werden für jeden Kanal unabhängig die Einsatzzeitpunkte geschätzt, wobei als Merkmal die relative Amplitudenveränderung zum vorherigen Fenster verwendet wird. Anschließend werden die Schätzungen aller Kanäle zusammengefasst, wobei jeder geschätzte Einsatzzeitpunkt als möglicher Kandidat betrachtet wird. Für jeden Kandidaten wird mittels eines Lautheitsmodells (nach Moore, Glasberg und Baer (1997)) ein Lautheitswert geschätzt. Alle Kandidaten, deren Lautheitswert einen globalen Schwellenwert unterschreiten, werden aussortiert. Abschließend werden noch Einsatzzeitpunkte, die innerhalb des Toleranzbereichs liegen (dort 50 ms), zusammengefasst, indem die Zeitpunkte mit dem niedrigeren Lautheitswerten aussortiert werden, bzw. bei gleichem Wert der Median gewählt wird. Die Ergebnisse der Studie fallen für verschiedene

Stücke sehr unterschiedlich aus (7% bis 95 % korrekt erkannte Toneinsätze) und sind schwierig mit anderen Studien zu vergleichen, da nur ein sehr kleiner Datensatz untersucht wird. Aus den Ergebnissen lässt sich jedoch schließen, dass das Verfahren für Musik mit hoher Dynamik versagt. Dafür verantwortlich ist vor allem der globale Schwellenwert, der Einsätze leiser Töne aussortiert. Aber auch die Verwendung des Lautheitsmodells ist für polyphone Musik problematisch, da so die Kandidatenauswahl von der Lautstärke anderer Töne abhängig ist.

In Holzapfel u. a. (2010) werden drei verschiedene Merkmale der Einsatzzeiterkennung zu einer Gesamtschätzung zusammengefasst. Neben *Spectral Flux* werden die Phasenverschiebung und die Grundfrequenzänderung benachbarter Fenster als Merkmale verwendet. Dabei wird für die Grundfrequenzschätzung der YIN-Algorithmus verwendet (De Cheveigné und Kawahara (2002), siehe Kapitel 4.1.1). Die Kombination der Merkmale wird damit motiviert, dass vermutlich auch bei der menschlichen Identifikation von Toneinsätzen Informationen über die spektrale Amplitude, die Phase und die Tonhöhe zusammen kombiniert werden. Es wird berichtet, dass für die Merkmalsaggregation weder eine Linearkombination noch ein Klassifikationsverfahren erfolgreich waren, ohne jedoch auf Details einzugehen. Allerdings wurden für das Klassifikationsverfahren, im Unterschied zu dem bereits beschriebenen Verfahren aus Bauer (2016), nur die aktuellen Werte der Merkmale verwendet. Holzapfel et al. berichten dann auch selbst, dass ein Hauptproblem ihres Klassifikationsverfahrens die unterschiedlichen Latenzen der verwendeten Merkmale sind: Phasenverschiebung erkennt Einsätze direkt am Anfang, während *Spectral Flux* die maximale Energieänderung erkennt, die meistens mit einer zeitlichen Verzögerung verbunden ist. Daher wird, wie in Klapuri (1999), auch in diesem Artikel die zweite Aggregationsvariante bevorzugt. Nachdem die drei binären Schätzvektoren vorliegen (für jedes Fenster: 1 = „Toneinsatz“, 0 = „kein Toneinsatz“), werden sie durch Summierung und Glättung zusammengefasst (wie genau wird in dem Artikel nicht beschrieben). Die Ergebnisse der Studie zeigen, dass dieses Verfahren zumindest für monophone Musik sehr gut funktioniert, denn der F-Wert wird gegenüber der besten Einzelschätzung (*Spectral Flux*) um 0.08 verbessert (0.82 gegenüber 0.74). Allerdings funktioniert das Verfahren für polyphone Musik nicht besser als die beste Einzelschätzung. Dies wird jedoch damit begründet, dass in diesem Fall die Grundfrequenzschätzung des YIN-Algorithmus sehr fehleranfällig ist. Eine Aggregation, die nur die anderen beiden Merkmale berücksichtigt, ergibt immerhin eine leichte Verbesserung (0.80 statt 0.78). Ein Nachteil des Verfahrens ist, dass die Einzelschätzungen auf gleichen Sprungweiten zwischen den Fenstern basieren sollten, da der Rechenzeitbedarf für Summierung und Glättung sonst stark steigt.

In dieser Arbeit werden drei Varianten für die Zusammenfassung der Einzelschätzungen der Kanäle getestet. Die erste Variante ist das einfache Verfahren, das nur den Kanal berücksichtigt, der in der Trainingsphase am erfolgreichsten abgeschnitten hat. Die zweite Variante ist die Quantilsaggregation aus Bauer u. a. (2014), wobei hier jedoch auch der Parameter p (welches Quantil) optimiert wird. Zudem wird natürlich auch hier der klassische Basisalgorithmus aus Kapitel 3.1 verwendet.

Die dritte Variante ist eine Aggregation der Einzelschätzungen, die ähnlich wie das Verfahren von Klapuri (1999) funktioniert. Im Gegensatz zu dessen Verfahren, werden die Kandidateneinsatzzeitpunkte jedoch nicht durch einen globalen Lautheitsschwellenwert bewertet, sondern ähnlich wie in Holzapfel u. a. (2010), durch die Anzahl von anderen Kandidaten in der unmittelbaren zeitlichen Umgebung. Dadurch ist das Verfahren nicht von der Laustärke abhängig und somit auch für dynamische Musik verwendbar. Im Unterschied zu Holzapfel u. a. (2010) werden hier jedoch auch asynchrone Schätzungen zugelassen, das heißt für jede Kanalschätzung dürfen alle Parameter – auch Fenstergröße und Sprungweite – unabhängig optimiert werden. Anstatt der Berechnung einer vollständigen Funktion, die jeden Abtastzeitpunkt bezüglich der Anzahl an Kandidaten in der Umgebung bewertet, werden hier in einer effizienteren Weise nur noch die Kandidatenzeitpunkte bewertet.

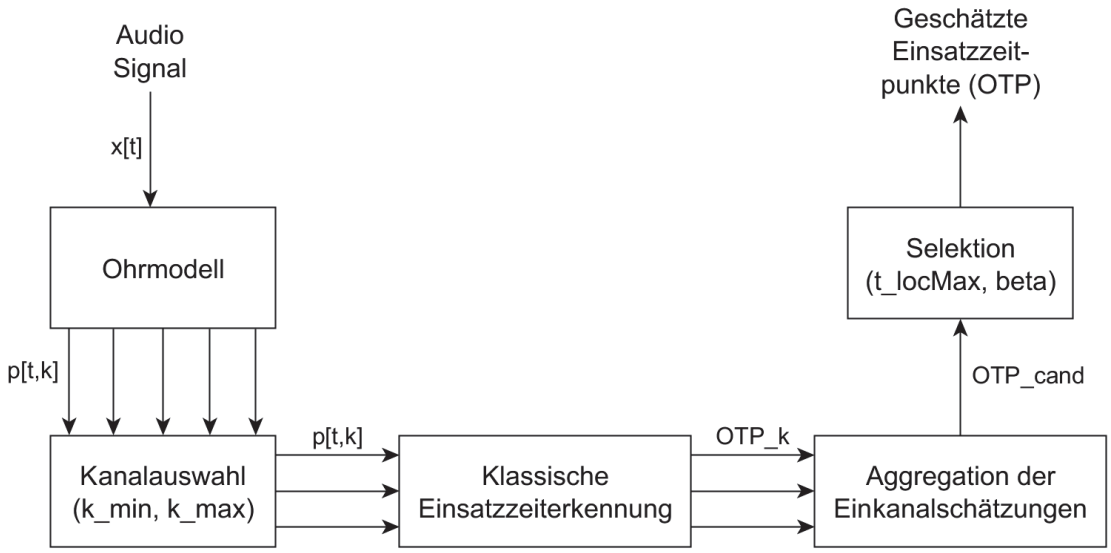

Abbildung 3.3: Blockdiagramm für die Einsatzzeiterkennung mit Ohrmodell (3. Variante).

Der Ansatz ist in Abbildung 3.3 skizziert. Auch hier sind die zu optimierenden Parameter

in Klammern angegeben. Da vor allem die tiefsten und höchsten Kanäle relativ schlechte Einzelschätzungen generieren (siehe Kapitel 7.1.1), wird das Verfahren noch durch eine Kanalselektion ergänzt. Dabei werden die zu berücksichtigenden Kanäle durch den minimalen $k_{\min}$ und maximalen $k_{\max}$ Kanal definiert. Alle geschätzten Einsatzzeitpunkte der übrig gebliebenen Kanäle werden zu einer Gesamtmenge von möglichen Kandidaten zusammengefasst:

$$\mathrm{OTP}_{\mathrm{cand}} = \bigcup_{k=k_{\min}}^{k_{\max}} \mathrm{OTP}_k \,. \tag{3.14}$$

In dieser Menge sind offensichtlich viele Kandidaten mehrmals enthalten, mit minimalen zeitlichen Verschiebungen. Diese müssen daher in einer geeigneten Weise zusammengefasst werden. Eine weitere Aufgabe ist es, falsche Kandidaten zu entfernen. Beide Aufgaben werden durch das folgende Verfahren gelöst. Für jeden Kandidaten wird die Anzahl an Schätzungen in seiner zeitlichen Nachbarschaft gemessen, wobei diese durch ± 25 ms definiert ist (entsprechend zum Toleranzbereich des F-Maßes). Anschließend werden nur solche Kandidaten akzeptiert, die ein lokales Maximum bilden und zudem einen globalen Schwellenwert übersteigen. Der Schwellenwert ist definiert durch

$$\beta \cdot (k_{\max} - k_{\min} + 1), \tag{3.15}$$

wobei β ein zu optimierender Parameter ist. Für jeden Kandidatenzeitpunkt n wird das Intervall, innerhalb dessen die Maximumsbedingung erfüllt sein muss auf $[n - t_{\mathrm{locMax}}, \ldots, n + t_{\mathrm{locMax}}]$ gesetzt, wobei t_{locMax} ein weiterer zu optimierender Parameter ist. Insgesamt ergeben sich durch dieses Verfahren vier zusätzliche Parameter, deren sinnvollen Wertebereiche in Tabelle 3.2 aufgelistet sind und die in einem zweiten MBO-Lauf optimiert werden. Da vier Parameter relativ schnell zu schätzen sind und auch die Auswertung eines einzelnen Punktes hier deutlich schneller geht als beim ersten MBO-Lauf – es müssen nur noch die geschätzten Zeitpunkte durchlaufen werden und nicht mehr die kompletten Musikausschnitte – ist der zusätzlich benötigte Rechenzeitbedarf vernachlässigbar klein.

Die Adaption hin zur predominanten Einsatzzeiterkennung wird auch für die Ohrmodellvariante durch eine Parameteranpassung mit MBO bezüglich der reduzierten Zielmenge an tatsächlichen Einsatzzeitpunkten erreicht.

Parametername	Minimaler Wert	Maximaler Wert
t_{locMax}	0 s	0.125 s
β	0	1
k_{min}	1	20
k_{max}	21	41

Tabelle 3.2: Parameter der Einsatzzeiterkennung mit Ohrmodell (3. Variante) für den zweiten MBO-Lauf.

4 Tonhöhenerkennung

Tonhöhenerkennung ($\widehat{=}$ Tonhöhenschätzung) wird oft als Synonym für Grundfrequenzschätzung verstanden. Jedoch ist Tonhöhe im eigentlichen Sinne ein perzeptives Maß, während die Grundfrequenz ein physikalisches Maß ist. In dieser Arbeit wird Tonhöhenerkennung als eine Grundfrequenzschätzung mit einer Fehlertoleranz von einem halben Halbton definiert. In der Frequenzskala entspricht dies einem Fehler von ca. 3%, was in etwa der Frequenzauflösung eines Normalhörenden entspricht. Zudem wird davon ausgegangen, dass die wahren Tonanfangs- und -endzeitpunkte bekannt sind, um eine Fehlerfortpflanzung der fehleranfälligen Einsatzzeiterkennung zu verhindern. Diese Information wird dazu verwendet, die Ohrmodellausgabe (bzw. das Originalsignal für das Vergleichsverfahren ohne Ohrmodell) in Melodietonsegmente zeitlich aufzuteilen. Die Aufgabe besteht somit darin, für jedes Segment die richtige Grundfrequenz zu schätzen. Die Güte dieser Schätzung wird durch die Fehlerrate, unter Berücksichtigung der Toleranz, definiert.

Die meisten Verfahren zur automatischen Tonhöhenerkennung basieren entweder auf der Autokorrelationsfunktion (ACF) oder sie führen eine Spektralanalyse im Frequenzbereich durch (z.B. mit der DFT), wobei potentielle Grundfrequenzen bezüglich ihrer entsprechenden Obertöne analysiert werden. Für beide Ansätze besteht das Hauptproblem darin, dass häufig mehrere Kandidaten in Frage kommen und es schwierig ist zu entscheiden, welcher davon die gesuchte Grundfrequenz ist. Bei polyphoner Musik ist dieses Problem auf Grund von möglichen Überlappungen der Obertöne zusätzlich erschwert. Die Autokorrelationsmethode zur Tonhöhenerkennung wird in Abschnitt 4.1 und ihre Erweiterung, der YIN-Algorithmus (De Cheveigné und Kawahara, 2002), in Abschnitt 4.1.1 erläutert.

Für die Tonhöhenerkennung existieren auch Ansätze, die ein auditorischen Modell oder zumindest dessen Teilkomponenten verwenden. Motivation dafür ist, neben dem Wunsch zur Nachbildung der menschlichen Wahrnehmung, auch die immer noch existierende Unterlegenheit der automatischen Verfahren, vor allem in Situationen mit mehreren Schallquellen. In McLeod (2009) wird ein Außen-/Mittelohr-Filter in der Vorverarbeitung

verwendet, wodurch die Anzahl der Oktavfehler gesenkt wird. Ein vollständiges auditorisches Modell wird in Meddis und Hewitt (1991) und in Meddis und O'Mard (1997) verwendet. Bei diesem Verfahren wird zunächst unabhängig für jeden Kanal die ACF berechnet, und anschließend werden diese Funktionen per Summierung zusammengefasst (SACF). Das SACF-Verfahren wird auch in dieser Arbeit experimentell untersucht und in Abschnitt 4.1.2 beschrieben.

Eine weitere Möglichkeit zur Berücksichtigung der ACFs mehrerer Kanäle, die vor allem für die polyphone Tonhöhenerkennung gedacht ist, wird in Klapuri (2008) vorgestellt. Dort wird die Stärke jeder möglichen Grundfrequenz (die Menge der lokalen Maxima der SACF) als eine gewichtete Summe der spektralen Amplituden der harmonischen Teiltöne kalkuliert. Die Erkennung weiterer Tonhöhen wird dann durch ein iteratives Verfahren gelöst, indem immer abwechselnd eine Tonhöhe geschätzt wird und anschließend die korrespondierenden Frequenzen (bezüglich der Teiltöne) abgeschwächt werden. Leider ist der Programmcode dieses Verfahrens aus Lizenzgründen nicht frei zugänglich, weshalb das Verfahren in den Experimenten dieser Arbeit nicht berücksichtigt wird.

Für die Grundfrequenzschätzung im spektralen Bereich wird in Duan, Pardo und Zhang (2010) ein Maximum-Likelihood Ansatz gewählt, um so das richtige Maximum der DFT zu identifizieren. Eine weitere Möglichkeit dafür ist statistische Klassifikation, die in Klapuri (2009) vorgeschlagen wird. Im Rahmen dieser Arbeit wurde ein klassifikationsbasiertes Verfahren entwickelt, bei dem jeder Kanal genau einen Grundfrequenzkandidaten vorschlägt, wodurch die Kanalnummer die Zielvariable wird. Für das Klassifikationsmodell waren ursprünglich nur spektralbasierte Merkmale der Ohrmodellausgabe vorgesehen, es können aber auch Merkmale verwendet werden, die aus der ACF extrahiert werden. Das Verfahren wird in Abschnitt 4.2 vorgestellt.

Bei der predominanten Tonhöhenerkennung sollen aus einem polyphonen Gemisch die Tonhöhen der Melodietöne erkannt werden. Alle bislang vorgestellten Verfahren können leicht dahingehend angepasst werden. Für die Klassifikations- und die Autokorrelationsmethoden erhöht sich hier lediglich die Schwierigkeit gegenüber der monophonen Variante. Weiterhin ist die dominierende Grundfrequenz bzw. Grundperiode gesucht, aber die Zahl der in Frage kommenden Kandidaten steigt. Auch iterative polyphone Verfahren, wie das von Klapuri (2008), können problemlos verwendet werden, indem lediglich die Grundfrequenz berücksichtigt wird, die als erstes gefunden wird.

Für die predominante Tonhöhenerkennung mit Ohrmodell werden für die Experimente in Kapitel 6 und 7) das SACF-Verfahren (mit zwei Methoden zur Peakauswahl), sowie das

entwickelte Klassifikationsverfahren (sowohl mit DFT- als auch mit ACF-Merkmalen) getestet. Als Vergleichsverfahren ohne Ohrmodell wird der anerkannte YIN-Algorithmus verwendet.

4.1 Autokorrelationsmethode

Die Autokorrelationsfunktion (ACF) eines Signals $x[t]$ ist durch

$$r[t,l] = \sum_{j=t+1}^{t+M} x[j]x[j+l] \tag{4.1}$$

definiert, wobei $r[t,l]$ die ACF mit Lag l zum Zeitpunkt t unter Berücksichtigung von M Abtastzeitpunkten ist. Da in dieser Arbeit auf Grund der Segmentierung $x[t]$ lediglich aus einem Melodieton besteht, ist hier prinzipiell nur $r[0,l]$ unter Berücksichtigung aller Abtastzeitpunkte von Interesse. Dennoch wird für alle Formeln in diesem Abschnitt die allgemeine Definition gewählt. Durch die ACF werden periodische Ähnlichkeiten hervorgehoben, denn ihr Funktionswert wird größer, je ähnlicher die beiden Folgen $x[j]$ und $x[j+l]$ sind. Der maximale Wert wird daher für $l = 0$ erreicht. Gemäß Gleichung 4.1 gilt in diesem Fall

$$r[t,0] = \sum_{j=t+1}^{t+M} x[j]^2, \tag{4.2}$$

was auch als die Energie des Signals bezeichnet wird (Rihaczek, 1968). Für vollkommen harmonische Signale wird der maximale Wert jedoch auch für die Periodendauer der Grundfrequenz ($T = 1/F_0$) erreicht, sowie für alle ganzzahligen Vielfache dieser Periodendauer. Diese entsprechen auf Grund des inversen Zusammenhangs zwischen Frequenz und Periode ganzzahligen Teilern der Grundfrequenz, sogenannten Untertönen. Bei komplexen Tönen gibt es zudem auch weitere lokale Maxima für die Perioden der Obertöne, die sogar den Funktionswert der Grundfrequenz übertreffen können. Hieran erkennt man bereits einige Probleme der Autokorrelationsmethode: Es ist oft nicht klar, welches Maximum die Periode der Grundfrequenz ist. Ein hoher Wert von $r[t,l]$ kann bedeuten, dass $1/l$ die Grundfrequenz ist, diese Frequenz kann aber auch ein Ober- oder Unterton sein. Weiterhin kann ein hoher Wert für niedrige Lags darauf hindeuten, dass eine entsprechende hohe Grundfrequenz vorliegt, aber der Grund dafür kann auch die Nachbarschaft zu $l = 0$ sein, möglicherweise noch durch einen Oberton oder hochfrequentes Rauschen überlagert.

Erweiterungen der Autokorrelationsmethode sind der YIN-Algorithmus und die SACF Methode, die in den folgenden beiden Abschnitten beschrieben werden.

4.1.1 YIN-Algorithmus

Der YIN-Algorithmus (De Cheveigné und Kawahara, 2002) kann als eine Verbesserung der Autokorrelationsfunktion angesehen werden, obwohl die erste Verbesserung darin besteht die Autokorrelation durch eine Differenzenfunktion zu ersetzen. Diese ist definiert durch

$$d[t,l] = \sum_{j=t+1}^{t+M} (x[j] - x[j+l])^2. \tag{4.3}$$

Im Gegensatz zur ACF wird hier nach Minima gesucht, wobei das globale Minimum $d[t,0] = 0$ ist. Man kann die Differenzenfunktion direkt aus der Autokorrelationsfunktion ableiten:

$$d[t,l] = r[t,0] + r[t+l,0] - 2r[t,l], \tag{4.4}$$

wobei $r[t,0]$ und $r[t+l,0]$ die Energie des Signals zum Zeitpunkt t bzw. $t+l$ messen (siehe Gleichung 4.2). Wären diese Werte konstant, wäre die Differenzenfunktion lediglich das Negative der Autokorrelationsfunktion (mit einer konstanten Verschiebung). Jedoch ist auch $r[t+l,0]$ von l abhängig, wodurch Perioden l bevorzugt werden, für die $r[t+l,0]$ niedrig ist. In den Experimenten in De Cheveigné und Kawahara (2002) verringert sich allein durch die Verwendung der Differenzenfunktion die Fehlerraten von 10% auf 1.95%. Die Autoren begründen dies damit, dass die Autokorrelationsfunktion Fehler bei Amplitudenänderungen macht. Denn bei einem höheren Lag werden gemäß Gleichung 4.1 mehr zukünftige Werte berücksichtigt, wodurch bei einem Amplitudenanstieg höhere Lags bevorzugt werden. Dieser Effekt wird bei der Differenzenfunktion durch den Energieterm $r[t+l,0]$ kompensiert.

Der zweite Verbesserungsschritt des YIN-Algorithmus ist die Ersetzung der Differenzenfunktion durch die kumulative mittlere normalisierte Differenzenfunktion, die durch

$$d'[t,l] = \begin{cases} 1, & \text{falls } l = 0 \\ {}^{d[t,l]}\!/_{\frac{1}{l}\sum_{j=1}^{l} d[t,j]}, & \text{sonst} \end{cases} \tag{4.5}$$

definiert ist. Durch diese Modifikation wird das Problem des Minimalwertes bei $l = 0$ behoben, ohne aber niedrige Perioden kategorisch auszuschließen. Denn $d'[t,l]$ startet

für $l = 0$ mit 1 anstatt mit 0, und für alle anderen Lags werden die ursprünglichen Werte im Verhältnis zum Mittelwert der ursprünglichen Werte aller niedrigeren Lags betrachtet. Hierdurch wird auch die Steigung der ursprünglichen Funktion berücksichtigt, so dass beispielsweise Lags nahe Null bestraft werden. In den Experimenten von Cheveigne und Kawahara bringt dieser Schritt immerhin eine kleine Verbesserung der Fehlerraten (1.69% statt 1.95%).

Der dritte Schritt zur besseren Grundfrequenzschätzung ist die Einführung eines globalen Schwellenwerts. Anstatt das Minimum von $d'[t, l]$ als Grundfrequenz zu schätzen wird stattdessen die kleinste Periode gewählt, die kleiner als ein vorher festgesetzter Schwellenwert ist und nur wenn keine solche Periode existiert, wird das globale Minimum gewählt. Dadurch wird das Problem behoben, dass bei der Differenzenfunktion (und wie oben bereits angemerkt auch bei der Autokorrelationsfunktion) häufig der Extremwert bei einer vielfachen Periode der Grundfrequenz zu finden ist, und dann fälschlicherweise die Frequenz eines Untertons gewählt wird. Für den Schwellenwert muss allerdings ein guter Kompromiss gefunden werden, denn diese Änderung kann natürlich auch zu einer zu starken Bevorzugung von kleinen Perioden führen. In den Experimenten von Cheveigne und Kawahara verbessert die Einführung eines Schwellenwerts von 0.1 die Fehlerraten nochmal erheblich (0.78% statt 1.69%).

Bei der Differenzenfunktion (wie auch bei der Autokorrelationsfunktion) kann es zu Rundungsfehlern von bis zu einer halben Abtastperiode kommen, wenn die zu schätzende Periode kein Vielfaches der Abtastperiode ist. Bei einer Abtastrate von 44100 (CD-Qualität) ist dieser Fehler allerdings verhältnismäßig klein. Falls beispielsweise statt einer korrekten Periode von 100.5 Abtastzeitpunkten die Periode 100 geschätzt wird, wird statt der korrekten 441 Hz eine Frequenz von 438.8 Hz vorhergesagt, womit der relative Fehler 0.5% ist, was deutlich unter der definierten Toleranzschwelle von 3% (ein halber Halbton) liegt. Allerdings wird dieser Fehler für höhere Frequenzen größer, so dass ab einer Frequenz ab etwa 2500 Hz der relative Fehler größer als 3% werden kann.[1] Es kann allerdings auch passieren, dass durch diesen Abtastfehler, der zu erreichende Schwellenwert knapp verfehlt wird und somit eine vollkommen falsche Tonhöhe geschätzt wird. Dieses Problem wird im YIN-Algorithmus durch eine parabolische Interpolation, dem vierten Verbesserungsschritt, gelöst. Dabei wird jedes lokale Minimum von $d'[t, l]$ und seine direkten Nachbarn an eine Parabel angepasst, wodurch das tatsächliche Minimum und der entsprechende Funktionswert exakter geschätzt werden. Diese Änderung brachte

[1] Dies sind allerdings Tonhöhen, die nur in wenigen Musikstücken vorkommen.

in den Experimenten von Cheveigne und Kawahara keine große Verbesserung (0.77% statt 0.78%), was allerdings damit begründet wird, dass die untersuchten Tonhöhen niedrig waren.

Der letzte Verbesserungsschritt des YIN-Algorithmus ist es t minimal zu variieren und dann das Minimum aus allen untersuchten Zeitpunkte auszuwählen. Dies wird damit begründet, dass Töne oft nicht vollkommen stationär sind und es stattdessen zeitliche Fluktuationen gibt, so dass die Tonhöhe möglicherweise bei einer leichten Verschiebung von t besser erkannt werden kann. Daher wird für die Untersuchung des Zeitpunkts t das Intervall $[t - {}^{T_{max}}/_2, t + {}^{T_{max}}/_2]$ untersucht, wobei T_{max} die maximal erwartete Periode ist. Durch diesen letzten Schritt wird in den Experimenten von Cheveigne und Kawahara die Fehlerrate auf 0.5% gesenkt (von 0.77%).

4.1.2 SACF-Verfahren

Für die Ohrmodellausgabe wird die ACF auf den Spikewahrscheinlichkeiten $p[t, k]$ durch

$$r_{\mathrm{AM}}[t, l, k] = \sum_{j=t+1}^{t+M} p[j, k]p[j + l, k] \tag{4.6}$$

definiert. Dabei ist $r_{\mathrm{AM}}[t, l, k]$ die ACF des Kanals k mit Lag l zum Zeitpunkt t unter Berücksichtigung von M Abtastzeitpunkten.

In Meddis und Hewitt (1991) wird vorgeschlagen, die ACFs der Kanäle per Summierung (Mittelwertbildung) zu einer summierten Autokorrelationsfunktion (SACF) zusammenzufügen. Durch eine Analyse der Maxima dieser Funktion kann dann wie bei der Autokorrelationsmethode die Grundfrequenz ermittelt werden. Die SACF $s[t, l]$ ist durch

$$s[t, l] = \frac{1}{K} \sum_{k=1}^{K} r_{\mathrm{AM}}[t, l, k] \tag{4.7}$$

definiert, wobei K die Anzahl der Kanäle ist. In Meddis und O'Mard (1997) wird gezeigt, dass die Ergebnisse des SACF Verfahrens, bezüglich einiger psychophysikalischer Phänomene, wie Tonhöhenerkennung bei fehlender Grundfrequenz, vergleichbar zur menschlichen Wahrnehmung sind. Für komplexe musikalische Signale ist das Verfahren jedoch noch nicht getestet.

Die Maxima der SACF sind Indikatoren für die wahrgenommene Tonhöhe. Auch bei diesem Verfahren besteht jedoch die Schwierigkeit den richtigen Peak auszuwählen. Die einfachste Methode ist die Suche nach dem global maximalen Wert:

$$\underset{l\in\{T_{\min},\ldots,T_{\max}\}}{\operatorname{argmax}} s[t,l], \tag{4.8}$$

wobei $T_{\min}$ und $T_{\max}$ die minimal bzw. maximal zu erwartende Grundperiode bezeichnen. Diese Methode identifiziert jedoch oft fälschlicherweise ganzzahlige Vielfache der Grundperiode. Um kleinere Perioden zu bevorzugen, bietet sich als Alternative das folgende Verfahren an:

$$\min\left[l \in l_{locMax} \colon s[t,l] > \lambda \cdot \max(s[t,l])\right], \tag{4.9}$$

wobei l_{locMax} die Menge der lokalen Maxima von s(t,l) mit $l \in \{T_{\min},\ldots,T_{\max}\}$ ist. Diese Methode wählt also den Peak mit der kleinsten Periode aus, der größer als der Schwellenwert $\lambda \in [0,1]$ ist. In den Experimenten dieser Arbeit wird λ bezüglich einer Trainingsmenge optimiert.

4.2 Klassifikationsmethode

In Weihs, Friedrichs und Bischl (2012) und Friedrichs und Weihs (2013) haben wir ein klassifikationsbasiertes Verfahren zur Identifikation aller Partialtöne von (monophonen) komplexen Tönen unter Verwendung eines Ohrmodells vorgestellt. Dieses Verfahren ist durch leichte Modifikationen auch für die Grundfrequenzschätzung anwendbar. Zunächst wird in Abschnitt 4.2.1 das ursprüngliche Verfahren zur Partialtonerkennung beschrieben. Anschließend folgt in Abschnitt 4.2.2 die Beschreibung des modifizierten Verfahrens zur Grundfrequenzschätzung.

4.2.1 Identifikation der Partialtöne

Die ursprüngliche Idee des Verfahrens war es, die Ohrmodellausgabe zu invertieren, um so den Höreindruck eines Schwerhörigen hörbar zu machen (Auralisierung). Zur Vereinfachung des Problems wurde die Musik dafür zunächst auf synthetische harmonische Töne beschränkt, wodurch das Problem auf die Schätzung der Teiltöne und ihrer Pegel reduziert wird. Später stellte sich jedoch heraus, dass dieser Ansatz nicht für reale polyphone Musik anwendbar ist. Allerdings kann das entwickelte Verfahren zur Bestimmung der Partialtöne

durch kleine Modifikationen auch zur Erkennung der Grundfrequenz verwendet werden, weshalb es in diesem Abschnitt beschrieben wird.

Die Grundidee des Verfahrens ist, dass jeder Kanal des Ohrmodells einen Partialton ($\widehat{=}$ Kandidat) vorschlägt, der anschließend durch ein binäres Klassifikationsverfahren entweder akzeptiert oder verworfen wird. Der Kandidat jedes Kanals ist dabei die Frequenz, bei der die DFT-Amplitude $|X[\mu]|$ maximal ist. Als einfache Schätzmethode wird die Frequenzlinie, also die Mittenfrequenz des entsprechenden Frequenzbandes, ausgegeben. Für die Klassifikation, die für jeden Kanal separat durchgeführt wird, werden 7 Merkmale verwendet: (1) Die Kandidatenfrequenz, (2) die spektrale Amplitude dieser Frequenz, (3) die Bandweite der Kandidatenfrequenz, (4,5) die spektralen Amplituden der Maxima links bzw. rechts der Kandidatenfrequenz, (6,7) die Distanz dieser Maxima zum Frequenzkandidaten.[2]

Für die Experimente wurde eine ältere Version des Ohrmodells von Meddis verwendet, das, im Unterschied zu dem in Kapitel 2.1 beschriebenen Modell, aus $K = 30$ Kanälen mit *Best Frequenzen* (BF) zwischen 100 Hz und 3 kHz besteht (statt 41 Kanälen mit BFs zwischen 100 und 6 kHz). Es wurden verschiedene Datensätze mit zufälligen synthetischen Tönen erzeugt, wobei drei Einflussgrößen variiert wurden: Die Tondauer (*dur*) mit den Ausprägungen 0.05 s, 0.2 s und 1.0 s, die Anzahl an Partialtönen (*tones*) mit den Ausprägungen 1,3,6 und 10 und die Wahrscheinlichkeit, mit der ein Teilton ausgelassen wird (*0prob*) mit den Ausprägungen 0.0, 0.2 und 0.5. Diese letzte Einflussgröße wurde aber nur bei 6 und 10 Partialtönen angewandt, bei 1 und 3 Partialtönen gilt immer $0prob = 0$. Insgesamt ergaben sich somit $(3 \cdot (2 + 2 \cdot 3) = 24$ mögliche Ausprägungskombination. Für jede Kombination wurde ein Datensatz mit 3000 Tönen erzeugt, deren Grundfrequenzen gleichverteilt auf die BFs der Kanäle aufgeteilt wurden. Für die anschließende Analyse wurde die Kanalnummer (*ch*) als zusätzlicher Einflussfaktor analysiert. Für jeden Datensatz und jeden Kanal wurde dann die Fehlklassifikationsrate des Klassifikationsmodells (Entscheidungsbaum) mittels zehnfacher Kreuzvalidierung evaluiert.

Die mittleren Fehlerraten, die eine Toleranz von einem halben Halbton berücksichtigen, sind im Anhang A – in Tabelle A.1 bezüglich der Datensätze und in Tabelle A.2 bezüglich der Kanäle – aufgelistet. Beim Vergleich der Datensätze fällt auf, dass die Fehlerraten mit der Komplexität der Töne ($\widehat{=}$ Anzahl der Partialtöne) steigen und dass die Töne mit sehr kurzer Tondauer am schlechtesten identifizierbar sind. Beim Vergleich der Kanäle fällt auf, dass die Kanäle mit hohen Nummern die größten Fehler aufweisen.

[2] Eine mathematische Beschreibung dieser Merkmale folgt in Abschnitt 4.2.2.

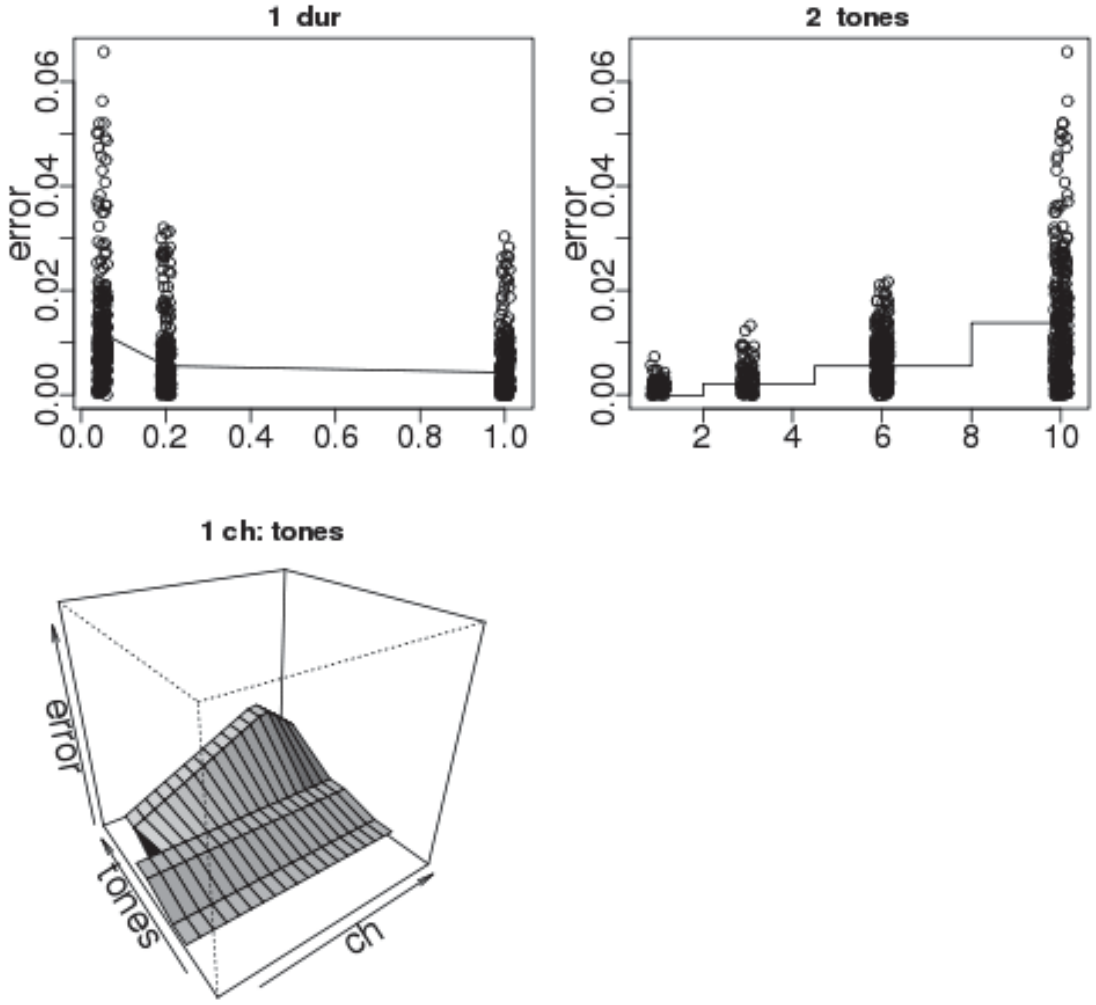

Abbildung 4.1: Effektplot des MARS-Modells für die Einflussfaktoren *ch*, *tones*, *dur* und *0prob* bezüglich der Fehlklassifikationsraten des Entscheidungsbaums. Es werden nur die wichtigsten Effekte aufgezeigt – $R^2 = 0.80$ (Weihs, Friedrichs und Bischl, 2012).

Abbildung 4.1 zeigt ein MARS (multiple adaptive regression splines) Modell (Friedman, Hastie und Tibshirani, 2001), das an diese Ergebnisse angepasst ist. Es zeigt, dass die Fehlerrate des Entscheidungsbaums im Wesentlichen von der Anzahl an Teiltönen und von der Tondauer abhängt. Der erste Effekt ist relativ klar der höheren Tonkomplexität zuzuordnen, die eine höhere Anzahl an Teiltönen mit sich bringt. Der zweite Effekt ist durch die ungenaue Frequenzschätzung der DFT für die kurzen Töne mit einer

Dauer von 0.05 Sekunden begründet. Für zukünftige Experimente könnte man die Schätzung des Maximalwerts von $|X[\mu]|$ durch ein Interpolationsverfahren ähnlich wie beim YIN-Algorithmus – z.B. durch das Quinn-Verfahren (Quinn, 1994) – exakter machen. Allerdings gilt auch für die menschliche Wahrnehmung, dass die Tonhöhenunterscheidung für Tonlängen unterhalb von 0.1 s ungenauer wird (Roederer und Mayer, 1999). Um diese Art des Fehlers auszuschließen, wird für das Versuchsdesign in Kapitel 6 eine minimale Tondauer von 0.1 s verwendet. Im unteren Plot des MARS-Modells ist zu erkennen, dass die Fehlerraten für höhere Kanalnummern größer sind, wenn der Ton aus vielen Partialtönen besteht. Dieser Effekt ist vermutlich dem Versuchsaufbau zuzuschreiben, der durch die gleichmäßige Aufteilung der Grundfrequenzen auf die Bandweiten der Kanäle dafür sorgt, dass bei den tiefen Kanälen nie ein Oberton klassifiziert werden muss, sondern nur Grundtöne. Dadurch ist die Klassifikationsaufgabe für diese Kanäle deutlich vereinfacht.

4.2.2 Identifikation der Grundfrequenz

Nachdem alle Teiltöne erkannt sind, ist eine einfache Grundfrequenzschätzung die Rückgabe des tiefsten Teiltons. Eine Erweiterung dieses Verfahrens ist eine Analyse, ob die anderen geschätzten Teiltöne auch zu dieser Grundfrequenz passen, und andernfalls den nächsthöheren Teilton zu wählen oder sogar eine vollkommen andere Frequenz (was z.B. bei einem Ton mit fehlendem Grundton Sinn macht). Bei polyphoner Musik wird es allerdings schwierig zu entscheiden, ob eine Grundfrequenz der Melodiestimme zuzuordnen ist. Dieses Problem könnte durch ein Klassifikationsmodell gelöst werde. Einfacher erscheint es dann aber, direkt ein Klassifikationsmodell zu erstellen, ohne den Umweg über die Partialtonerkennung zu nehmen. Dieses Modell betrachtet die extrahierten Merkmale aller Kanäle gemeinsam und identifiziert den Kanal – d.h. den richtigen Frequenzbereich –, der den (hoffentlich) richtigen Kandidaten vorschlägt. Im Folgenden werden die im vorherigen Unterkapitel beschriebenen Merkmale formal definiert und noch um 22 weitere Merkmale ergänzt.

Zunächst wird für jeden Kanal k die DFT (siehe Gleichung 2.7) $P[\mu, k]$, der Ohrmodellausgabe $p[t, k]$, mit $t = 1, \ldots, M$, berechnet. Auf Grund der Symmetrieeigenschaft der DFT müssen nur die ersten $M/2$ Fourier-Koeffizienten berücksichtigt werden. Anschließend bestimmt jeder Kanal einen Grundfrequenzkandidaten. Dafür wird die Frequenzlinie mit maximaler DFT-Amplitude gewählt, die innerhalb eines Frequenzbereichs um die *Best Frequenz* des Kanals liegt. Dieser Bereich ist durch die *Best Frequenzen* der beiden

benachbarten Kanäle begrenzt:

$$\mu^*[k] = \underset{\mu \in \{\mathrm{BF}[k-1],\mathrm{BF}[k-1]+1,\ldots,\mathrm{BF}[k+1]\}}{\arg\max} |P[\mu, k]|, \quad k = 1, \ldots, K, \tag{4.10}$$

wobei BF$[k]$ das Frequenzband bezeichnet, welche die *Best Frequenz* des Kanals k beinhaltet (für $k = 1, \ldots, K$). Diese liegt zwischen 100 Hz für den ersten und 6 kHz für den letzten (41.) Kanal (siehe Gleichung 2.1 in Kapitel 2.1). Für die Begrenzungen der Frequenzbereiche (innerhalb derer der Kandidat gesucht wird) des ersten und des letzten Kanals müssen zusätzlich $BF[0]$ und $BF[K+1]$ definiert werden. $BF[0]$ wird definiert als das Frequenzband, das die Frequenz 50 Hz enthält, und $BF[K+1]$ als das Frequenzband, das die Frequenz 10 kHz beinhaltet. Die Mittenfrequenz $CF(\mu)$ des Frequenzbandes $\mu^*[k]$ ist die Kandidatenfrequenz $c[k] = CF(\mu^*[k])$. Anstatt die Mittenfrequenz zu wählen, könnte eine Interpolationsmethode, wie das Quinn-Verfahren, das Verfahren etwas präziser machen. Auf Grund der zugelassenen Fehlertoleranz von einem halben Halbton ist der daraus resultierende Fehler jedoch für die Experimente in dieser Arbeit vernachlässigbar klein.

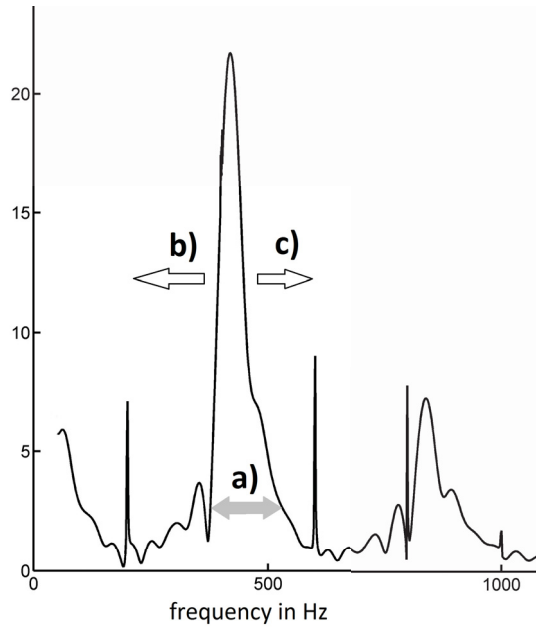

Abbildung 4.2: Drei Merkmale der Tonhöhenklassifikation: a) Bandweite $b[k]$ des Kandidaten, b) Abstand zum linken Maximum $d_{\mathrm{left}}[k]$ und c) Abstand zum rechten Maximum $d_{\mathrm{right}}[k]$ (Friedrichs und Weihs, 2013).

Das Klassifikationsziel ist die Identifikation des Kanals, dessen *Best Frequenz* einen möglichst geringen Abstand zur Grundfrequenz hat. Der Frequenzkandidat dieses Kanals wird dann als die geschätzte Grundfrequenz ausgegeben. Für die Erstellung des Klassifikationsmodells, werden die folgenden Merkmale verwendet, die für jeden Kanal (bzw. jeden Kandidaten) k unabhängig extrahiert werden.

- Die Kandidatenfrequenz $c[k]$.

- Die spektrale Amplitude des Kandidaten: $a_c[k] = |P[\mu^*[k], k]|$.

- Die Bandweite $b[k]$ des Kandidatenpeaks, definiert durch die Distanz zwischen den beiden nächsten Frequenzlinien zur linken bzw. rechten Seite, deren spektralen Amplituden geringer sind als 10% der Kandidatenamplitude $a_c[k]$ (siehe auch Abbildung 4.2):
$$b[k] = CF(\mu^*_{right}[k]) - CF(\mu^*_{left}[k]). \tag{4.11}$$
Die Ränder des Bandes sind definiert durch
$$\mu^*_{\text{right}}[k] = \min(\mu \in \{\mu^*[k], ..., M/2\} : a_c[k]/10 > |P[\mu, k]|), \tag{4.12}$$
wobei $\mu^*_{\text{right}}[k]$ auf $M/2$ gesetzt wird, falls kein solches μ existiert, und
$$\mu^*_{\text{left}}[k] = \max(\mu \in \{1, ..., \mu^*[k]\} : a_c[k]/10 > |P[\mu, k]|), \tag{4.13}$$
wobei $\mu^*_{\text{left}}[k]$ auf 1 gesetzt wird, falls kein solches μ existiert.

- Die Distanz des Frequenzkandidaten zu den beiden Maxima auf der linken bzw. rechten Seite des Kandidaten, beschränkt durch dessen Bandränder (zwei Merkmale: $d_{\text{left}}[k]$ und $d_{\text{right}}[k]$, siehe auch Abbildung 4.2):
$$d_{\text{left}}[k] = c[k] - CF(\operatorname*{argmax}_{\mu \in \{1 ... \mu^*_{left}[k]\}} (P[\mu, k])) \quad \text{und} \tag{4.14}$$
$$d_{\text{right}}[k] = CF(\operatorname*{argmax}_{\mu \in \{\mu^*_{right}[k] ... M/2\}} (P[\mu, k])) - c[k]. \tag{4.15}$$

- Die spektrale Amplitude dieser beiden Maxima (zwei Merkmale): $|P[\max_{left}[k]]|$ und $|P[\max_{right}[k]]|$.

Nr.	Merkmalsname	Mathematische Bezeichnung
1	Kandidatenfrequenz	$c[k]$
2	Kandidatenamplitude	$a_c[k]$
3	Kandidatenbandweite	$b[k]$
4	Distanz zum linken Maximum	$d_{\text{left}}[k]$
5	Distanz zum rechten Maximum	$d_{\text{right}}[k]$
6	Amplitude des linken Maximums	$\lvert P[\max_{left}[k]]\rvert$
7	Amplitude des rechten Maximums	$\lvert P[\max_{right}[k]]\rvert$
8	mittlere Feuerrate des Kanals	$p_{mean}[k]$
9	maximale Feuerrate des Kanals	$p_{\max}[k]$
10 - 18	mittlere Amplituden der 9 Partialtöne	$P_{\text{pl}}^{\text{mean}}[k]$
19 - 27	maximale Amplituden der 9 Partialtöne	$P_{\text{pl}}^{\max}[k]$
28	mittlere Amplitude des 1. Untertons	$P_{1/2}^{mean}[k]$
29	maximale Amplitude des 1. Untertons	$P_{1/2}^{max}[k]$

Tabelle 4.1: Verwendete Merkmale für die Tonhöhenklassifikation (für jeden Kanal k).

- Die mittlere und die maximale Spikewahrscheinlichkeit des Kanals:

$$p_{\text{mean}}[k] = \frac{1}{M}\sum_{t=1}^{M} p[t,k] \quad \text{und} \tag{4.16}$$

$$p_{\max}[k] = \max_{t\in\{1,\ldots,M\}}(p[t,k]). \tag{4.17}$$

- Die mittleren und die maximalen spektralen Amplituden der ersten neun Partialtöne ($\text{pl} = 1,\ldots,9$) über alle Kanäle:

$$P_{\text{pl}}^{\text{mean}}[k] = \frac{1}{K}\sum_{n=1}^{K} P[a(\text{pl}\cdot c[k]), n], \tag{4.18}$$

  wobei $a(i)$ das Frequenzband ist, das die Frequenz i beinhaltet, und

$$P_{\text{pl}}^{\max}[k] = \max_{n\in\{1,\ldots,K\}}(P[a(\text{pl}\cdot c[k]), n]). \tag{4.19}$$

- Gleichermaßen auch die mittlere und die maximale spektrale Amplitude des ersten Untertons (halbe Grundfrequenz) des Kandidaten, summiert über alle Kanäle: $P_{1/2}^{mean}[k]$ und $P_{1/2}^{max}[k]$.

Pro Kanal ergeben sich somit 29 Merkmale, was bei Verwendung des Ohrmodells ohne

Hörschädigung insgesamt $29 \cdot 41 = 1189$ Merkmale bedeutet. Diese Merkmale sind in Tabelle 4.1 zusammengefasst. Wie bereits erwähnt, ist das Verfahren nicht auf DFT-Merkmale beschränkt. Als Alternative wird deshalb in den Experimenten (Kapitel 6 und 7) auch ein Verfahren untersucht, dass alle DFT-Merkmale aus Tabelle 4.1 durch äquivalente Merkmale aus der ACF $r_{\mathrm{AM}}[t, l, k]$ eines Kanals extrahiert. Hierbei bleiben lediglich die im Zeitbereich definierten Merkmale $p_{mean}[k]$ und $p_{\max}[k]$ identisch.

5 Instrumentenerkennung

Ziel der Instrumentenerkennung ist die Identifikation der Instrumentenbesetzung eines Musikstücks oder einer Teilsequenz. Die Klang verschiedener Instrumente unterscheidet sich auf Grund einer unterschiedlichen Verteilung der Obertonintensitäten. Beispielsweise sind bei der Klarinette die Intensitäten fast vollständig auf die ungeraden Partialtöne verteilt, während die geraden Partialtöne nur schwach vorkommen. Allerdings ist die Obertonverteilung auch von anderen Eigenschaften abhängig, z.B. von der Tonhöhe, der Raumakustik und der Spielweise (Sandrock, 2013). Hinzu kommen zudem Unterschiede zwischen verschiedenen Repräsentanten des gleichen Musikinstruments. Neben der Obertonverteilung gibt es auch noch weitere Charakteristiken, die den Klang eines Instruments ausmachen, wie z.B. unharmonische Bestandteile oder die Dauer der Anschlagsphase. Da es keine allgemeingültigen intuitiven Regeln für die Unterscheidung von Musikinstrumenten gibt, wird Instrumentenerkennung üblicherweise durch überwachte Klassifikation, unter Verwendung einer großen Anzahl von Merkmalen, gelöst. Instrumentenerkennung wird im Folgenden daher auch als Instrumentenklassifikation bezeichnet.

Die typische Prozessabfolge der Instrumentenerkennung ist in Abbildung 5.1 illustriert. Zunächst wird eine geeignete Datenmenge von Musikbeispielen benötigt, die mit den zu klassifizierenden Instrumenten gelabelt sind. Die Art der Daten ist abhängig von der konkreten Anwendung und kann sehr unterschiedlich sein. Beispielsweise kann eine Beobachtungen durch einen Einzelton, aber auch durch ein komplettes Musikstück definiert sein. Die Art der Daten lässt sich durch vier Dimensionen beschreiben, die die Komplexität der spezifischen Anwendung definieren. Diese Dimensionen werden in Abschnitt 5.1 diskutiert.

Im nächsten Schritt wird die Taxonomie ($\widehat{=}$ Klassifikationsschema) definiert. Die einfachste Variante ist eine flache Taxonomie, bei der jeder Beobachtung direkt ein Instrumentenlabel zugeordnet wird. Da die Klangunterschiede verschiedener Instrumente unterschiedlich stark sind, kommt jedoch auch die Verwendung einer hierarchischen Taxonomie in Betracht. Diese ordnet beispielsweise eine Beobachtung zunächst einer Instrumentenfamilie

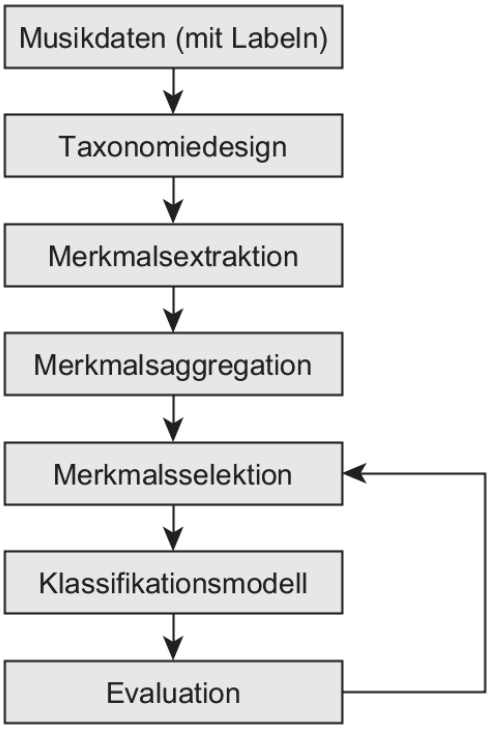

Abbildung 5.1: Schritte der Instrumentenerkennung.

zu, und erst anschließend werden die Beobachtungen dieser Familie durch ein zweites Klassifikationsmodell in die exakten Instrumente aufgeteilt. In Abschnitt 5.2 werden verschiedene Taxonomien beschrieben.

Für die Generierung eines Klassifikationsmodells muss zunächst die relevante Information durch Merkmale extrahiert werden. Welche Merkmale für die Instrumentenklassifikation in Betracht kommen, wird in Abschnitt 5.3 erörtert. Einige Merkmale werden nicht aus dem vollständigen Segment, sondern nur aus einem kleinen Teilbereich (Fenster oder auch *Frame* genannt) extrahiert. Diese Merkmale werden in jedem Fenster extrahiert, so dass anschließend eine Vielzahl von Ausprägungen des Merkmals vorliegt. Allein um die Merkmalsdimension zu reduzieren, ist es sinnvoll diese Ausprägungen in einem nächsten Schritt zusammenzufassen. Dieser Schritt wird in Abschnitt 5.3.3 beschrieben. Eine weitere Reduzierung der Merkmalsdimension wird durch eine Merkmalsselektion erreicht, die bereits in Kapitel 2.4 beschrieben ist. Der nächste Schritt ist die Wahl des Klassifikationsverfahrens und eventueller Hyperparameter (siehe Kapitel 2.3). Um die Güte eines Klassifikationsmodells zu bewerten und verschiedene Verfahren miteinander

vergleichen zu können, ist zum Schluss noch eine Evaluation nötig, durch die das Klassifikationsmodell durch vorher nicht gesehenen Beobachtungen getestet wird. Welches Evaluationsverfahren am geeignetsten ist, hängt von der Struktur der Daten, insbesondere von deren Größe, ab. In dieser Arbeit wird bei Vorstudien mit kleineren Datensätzen eine zehnfache Kreuzvalidierung verwendet und für die umfangreichen Experimente in den Kapiteln 6, 7 und 8 eine dreifache Kreuzvalidierung.

Die Variationsmöglichkeiten der beschriebenen Schritte wurden teilweise bereits in Vorstudien untersucht. Diese Studien werden in Abschnitt 5.4 kurz zusammengefasst.

5.1 Typen der Instrumentenerkennung

Die Komplexität der Daten für die Instrumentenerkennung kann durch vier Dimensionen beschrieben werden:

1. Zeitliche Struktur der Beobachtungen,
2. Spektrale Struktur der Beobachtungen,
3. Interklassenabstand,
4. Intraklassenabstand.

Von diesen Dimensionen hängt der Schwierigkeitsgrad der Klassifikationsentscheidung ab. Dies ist bei einem Vergleich der Ergebnisse verschiedener Studien zur Instrumentenklassifikation zu berücksichtigen. Im Folgenden werden die Dimensionen näher erläutert.

5.1.1 Zeitliche Struktur der Beobachtungen

Je nachdem welche Zeitabschnitte klassifiziert werden sollen, ist eine Beobachtung für die Instrumentenerkennung sehr unterschiedlich definiert. Beispielsweise kann eine Beobachtung ein komplettes Musikstück, ein einzelner Ton (bzw. Intervall oder Akkord) oder auch ein noch kürzeres zeitliches Segment sein. Auf Grund der geringeren Information ist die Klassifikation eines Datensatzes schwieriger, je kürzer die den Beobachtungen zu Grunde liegenden Zeitintervalle sind. Beispielsweise kann die Instrumentenidentifikation kompletter monophoner Musikstücke auf die Klassifikation kurzer Segmente mit anschließender Majoritätsentscheidung reduziert werden. Dieses Verfahren hat offensichtlich einen kleineren erwarteten Fehler als die Einzelentscheidungen für die Segmente. Weiterhin

erleichtert die Kenntnis über die Anfangs- und Endzeitpunkte der Töne die Erkennung, da hierdurch die zeitliche Struktur des Tonaufbaus berücksichtigt werden kann.

In den Experimenten dieser Arbeit wird die Instrumentenerkennung für jeden Ton unabhängig durchgeführt. Dafür wird, wie bei der Tonhöhenerkennung, die Ohrmodellausgabe in zeitliche Segmente entsprechend der wahren Anfangs- und Endzeiten der Melodietöne aufgeteilt. Jedes Melodietonsegment entspricht dann einer Beobachtung, die unabhängig geschätzt wird. Da innerhalb der Musikstücke das Melodieinstrument nicht wechselt, wären die Ergebnisse leicht durch eine gemeinsame Schätzung aller Töne zu verbessern (z.B. Majoritätsentscheidung). Für diese Arbeit stellt die Instrumentenerkennung jedoch keinen Selbstzweck dar, sondern sie dient der Schätzung, ob die Klangfarbe konstant gut wahrgenommen wird. Dies wird nur durch eine unabhängige Schätzung der Beobachtungen ermöglicht. Andernfalls würde es ausreichen einen Zeitpunkt zu ermitteln, zu dem die Begleitung pausiert, und nur diesen zu klassifizieren. Die Schätzung der Klangfarbenwahrnehmung bei polyphoner Musik wäre somit stark verfälscht.

5.1.2 Spektrale Struktur der Beobachtungen

Monophone Instrumentenerkennung, bei der in jedem Zeitpunkt nur ein Instrument spielt, ist deutlich einfacher als die polyphone Variante, in der mehrere Instrumente gleichzeitig spielen und somit die spektrale Struktur komplexer wird. Die meisten neueren Studien beschäftigen sich mit der polyphonen Variante. Dabei ergibt sich das Problem von überlappenden Obertönen der verschiedenen Töne, was einen nichtlinearen Effekt auf die Merkmalswerte hat. Beispielsweise können leisere Töne komplett von lauteren überdeckt werden. Eine spezielle Variante ist die predominante Instrumentenerkennung. Dieses Problem kann ähnlich wie die monophone Variante gelöst werden, wobei aber die höhere Komplexität der polyphonen Variante bestehen bleibt (Wieczorkowska, Kubera und Kubik-Komar, 2011). Die Problemkomplexität wird noch größer, wenn auch die Begleitinstrumente identifiziert werden sollen. Hierbei entsteht das Problem, wie korrelierte Ereignisse, die gleichzeitig stattfinden, klassifiziert werden können. Der naive Ansatz ist, eine Klasse für jede Instrumentenkombination zu generieren. Dies führt aber schnell zu einer extrem großen Anzahl von Klassen, wobei für einige Klassen nur sehr wenige oder sogar gar keine Beobachtung vorliegen. Eine Alternative ist, mit einer Quellentrennung zu starten, um anschließend jede Quelle unabhängig zu schätzen. Bei diesem Konzept werden allerdings Fehler der Quellentrennung, die in ihrer allgemeinen Form noch lange

nicht gelöst ist, fortgepflanzt. Eine dritte Möglichkeit ist eine Multilabel-Klassifikation, bei der jede Beobachtung mehreren Klassen zugeordnet werden kann (Sandrock, 2013).

In dieser Arbeit wird hauptsächlich die predominante Instrumentenerkennung untersucht, sowie für Vergleichszwecke auch die monophone Variante ohne Begleitung.

5.1.3 Interklassenabstand

Der Interklassenabstand der Instrumentenerkennung wird durch die Ähnlichkeit und die Anzahl der zu trennenden Instrumente beeinflusst. Eine große Ähnlichkeit und eine hohe Anzahl verschiedener Instrumente erhöhen die Komplexität. Andererseits wird die Klassifikationsentscheidung vereinfacht, wenn nur die Instrumentenfamilie zu identifizieren ist.

In den Experimenten der Kapitel 6, 7 und 8 werden (in erster Linie aus Rechenzeitgründen) nur drei Instrumente untersucht, die zudem relativ verschieden sind: Cello, Klarinette und Trompete.

5.1.4 Intraklassenabstand

Der Intraklassenabstand der Instrumentenerkennung ist abhängig von den untersuchten Musikinstrumenten und deren Repräsentanten. In einigen Anwendungen wird nur ein spezifischer Repräsentant eines Musikinstruments verwendet, wohingegen in anderen ein universelles Modell angestrebt wird, das für alle Repräsentanten des Instruments gültig ist. In diesem Fall wird die Streuung zwischen Beobachtungen der gleichen Klasse größer. Diese hängt auch davon ab, wie unterschiedlich die Bau- und Spielweisen der betrachteten Repräsentanten sind. Ein praktisches Problem ist allerdings, dass es keinen veröffentlichten Datensatz gibt, der umfangreich genug ist, ein universelles Modell trainieren und evaluieren zu können. Für den Intraklassenabstand ist zudem auch noch entscheidend, wie stark sich die Spektren der betrachteten Instrumente bei verschiedenen Tonhöhen unterscheiden. Beispielsweise lässt bei der Klarinette die übliche Dominanz der ungeraden Partialtöne bei hohen Tönen stark nach (Roederer und Mayer, 1999).

In dieser Arbeit wird die Instrumentenerkennung dazu verwendet, die Schärfe der Klangfarbenwahrnehmung für eine Hörschädigung zu schätzen. Ziel ist somit nicht die Erstellung eines universelles Modells zur Instrumentenerkennung. Daher wird pro Instrument nur ein Repräsentant verwendet.

5.2 Taxonomiedesign

Die einfachste Taxonomie für die Instrumentenerkennung ist die flache Variante, in der jede Beobachtung direkt einem Instrument zugeordnet wird. Ein Beispiel ist in Abbildung 5.2 zu sehen. Da die Problemkomplexität steigt, wenn die betrachteten Instrumente eine ähnliche Klangfarbe besitzen, wie z.B. Horn und Trompete, kommt auch eine hierarchische Taxonomie in Betracht. Diese ist vor allem für Anwendungen sinnvoll, bei denen die Kosten für verschiedene Fehlklassifikationen unterschiedlich hoch sind, beispielsweise weil die Erkennung der richtigen Instrumentenfamilie ein Teilziel ist. Es gibt zwei unterschiedliche Herangehensweisen zur Erstellung einer hierarchischen Taxonomie. Sie kann vordefiniert werden, z.B. durch eine Zusammenfassung von Instrumenten zu natürliche Instrumentengruppen, oder sie kann automatisch erzeugt werden, z.B. mit Hilfe eines Cluster-Verfahrens. Hierarchische Taxonomien können als Baum oder gerichteter Graph mit einer Klassifikationsentscheidung in jedem inneren Knoten visualisiert werden. Da jedes Klassifikationsmodell unabhängig gelernt wird, können in jedem Knoten andere Merkmale und andere Klassifikationsverfahren für die Klassifikationsentscheidung gewählt werden. In Vatolkin u. a. (2012) wurde gezeigt, dass die Menge der optimalen Merkmale für die Trennung verschiedener Instrumentengruppen sehr unterschiedlich ist. Prinzipiell könnten auch Merkmalsextraktion und Merkmalsaggregation in jedem Knoten unterschiedlich definiert werden, das würde aber einen hohen Mehraufwand bedeuten und ist deshalb in der Praxis unüblich.

Das Training einer hierarchischen Klassifikation läuft wie folgt ab. Zunächst wird ein Klassifikationsmodell für den obersten Knoten der Taxonomie unter Berücksichtigung aller Beobachtungen der Trainingsmenge erstellt. Anschließend werden für die Knoten der nächsten Ebene Klassifikationsmodelle gebildet, wobei aber in der Trainingsphase nur die Beobachtungen berücksichtigt werden, die in den Knoten einsortiert werden sollen. Dieses Vorgehen wird iterativ wiederholt bis die unterste Ebene erreicht ist. Ein Nachteil der hierarchischen Klassifikation ist, dass Fehler in höheren Ebenen zu allen darunterliegenden Ebenen weitergereicht werden.

Eine natürliche hierarchische Taxonomie ist die Einteilung in die bekannten Instrumentenfamilien: Schlaginstrumente, Streichinstrumente, Holzblasinstrumente und Blechblasinstrumente (siehe Abbildung 5.3). Diese Taxonomie hat allerdings den Nachteil, dass nicht alle Instrumente eindeutig zu einer der Gruppen gehören, denn zum Beispiel könnte man das Klavier zu den Schlag- aber auch zu den Streichinstrumenten zählen (Sandrock, 2013). Eine Taxonomie, bei der die Einordnung eindeutig ist, ist die

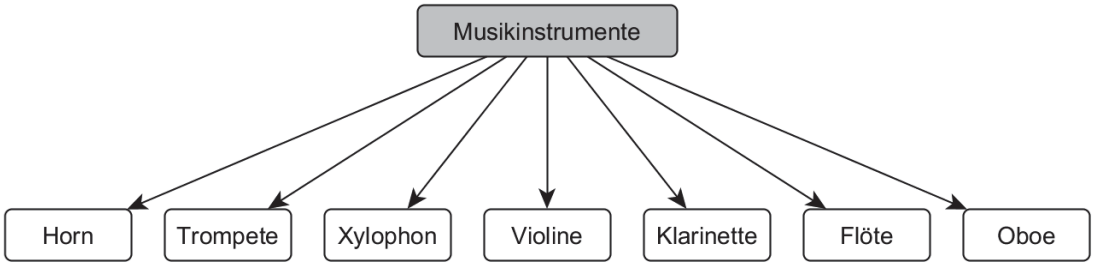

Abbildung 5.2: Flache Taxonomie.

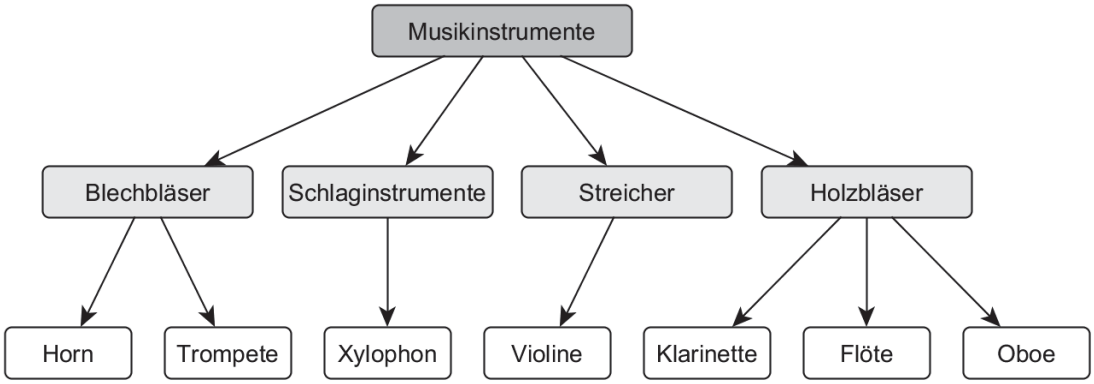

Abbildung 5.3: Hierarchische Taxonomie nach Instrumentenfamilie.

Hornbostel-Sachs-Systematik, in der die Instrumente entsprechend ihrer Tonerzeugung eingeordnet werden (Hornbostel und Sachs, 1961). Es besteht aus über 300 Kategorien, die über mehrere Ebenen angeordnet werden. Ein kleiner Ausschnitt des kompletten Systems ist in Abbildung 5.4 illustriert. Auf der ersten Ebene werden Instrumente in fünf Hauptkategorien eingeordnet: Idiophone, Membranophone, Chordophone, Aerophone und Elektrophone. Idiophone sind alle Instrumente, deren Körper direkt für die Klangerzeugung zuständig ist, ohne dass eine Membran oder Saite dafür benötigt wird. Dies umfasst alle Schlaginstrumente mit Ausnahme von Trommeln. Membranophone sind alle Instrumente, bei denen der Klang durch straff gespannte Membrane entsteht, wozu fast alle Arten von Trommeln zählen. Chordophone sind alle Instrumente bei denen eine oder mehrere Saiten zwischen fixierten Punkten gedehnt werden. Dies umfasst alle Streichinstrumente und das Klavier. Der Klang von Aerophonen wird durch Vibration der Luft erzeugt, wie in fast allen Blasinstrumenten. Elektrophone sind alle Instrumente, bei denen Elektrizität für die Klangerzeugung benötigt wird, wie z.B. Synthesizer oder Theremins.

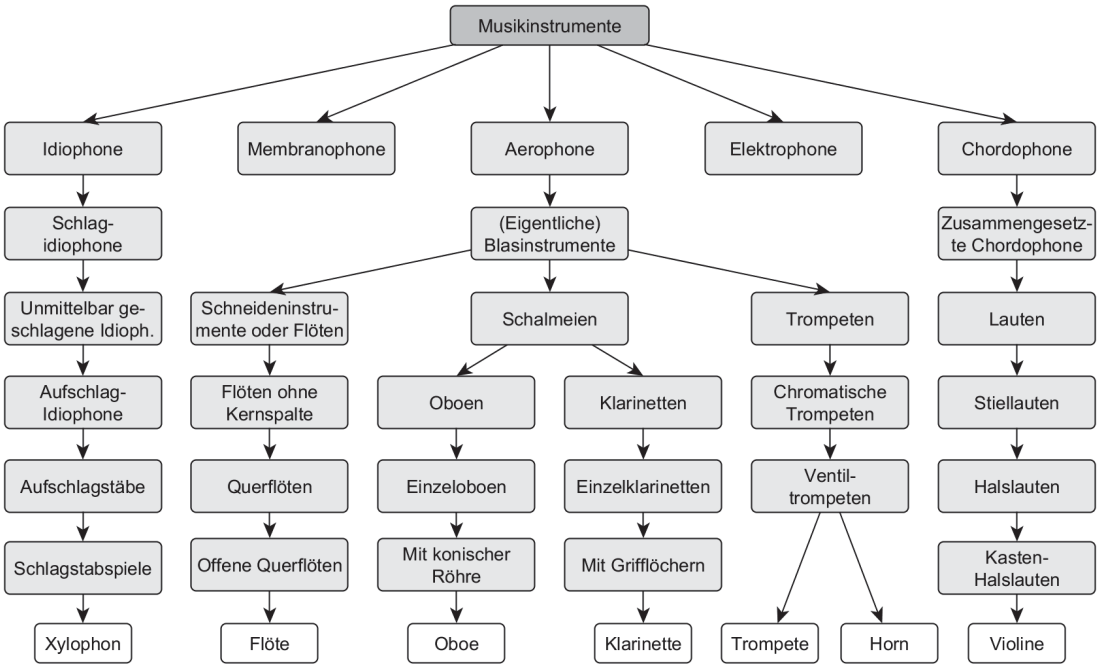

Abbildung 5.4: Hornbostel-Sachs-Taxonomie (Ausschnitt).

Wann eine hierarchische Klassifikation gegenüber einer flachen zu bevorzugen ist, ist nicht abschließend geklärt. Es dürfte aber datenabhängig sein, denn die Ergebnisse verschiedener Studien sind diesbezüglich unterschiedlich. In Essid, Richard und David (2006) wird eine automatische Taxonomie durch hierarchisch-agglomeratives Clustern generiert, wobei Klassen entsprechend eines geeigneten Ähnlichkeitsmaßes automatisch zusammengefasst werden. Die Autoren argumentieren, dass hierfür die euklidische Distanz nicht geeignet ist und testen stattdessen zwei probabilistische Distanzmaße. In ihren Ergebnissen, unter Verwendung einer SVM mit einem Gauß-Kern, erreicht ihr Ansatz leicht bessere Ergebnisse als die flache Taxonomie (36% gegenüber 39% Fehlklassifikation). Andererseits berichten Eronen und Klapuri (2000), dass sie keinen Nachweis für die Überlegenheit einer hierarchischen Klassifikation finden konnten.

Für die Experimente in den Kapiteln 6, 7 und 8 werden nur drei völlig unterschiedliche Instrumente betrachtet (Cello, Klarinette, Violine). Eine hierarchische Taxonomie macht daher wenig Sinn und stattdessen wird eine flache verwendet. Für zukünftige umfangreichere Experimente könnte jedoch eine hierarchische Taxonomie in Betracht kommen. Daher werden in einer Vorstudie in Abschnitt 5.4.4 die Taxonomien anhand eines Datensatzes mit 5 Instrumenten verglichen.

5.3 Merkmale

Am verbreitetsten für die Instrumentenklassifikation sind Klangfarbenmerkmale, die direkt aus der akustischen Wellenform extrahiert werden. Allerdings wurden bereits in Martin und Kim (1998) Klangfarbenmerkmale aus der Ohrmodellausgabe extrahiert, die allerdings bislang wenig Beachtung gefunden haben. Ein ganz neuer Ansatz sind biometrisch motivierte spektral-temporale Merkmale, die jedoch ein Modell der höheren zentralen auditorischen Ebenen benötigen und auf die daher hier nicht genauer eingegangen wird. In Patil und Elhilali (2015) wurden sie erfolgreich für Solomusik getestet. Im Folgenden Abschnitt 5.3.1 werden eine Reihe von herkömmlichen Merkmalen definiert und in Abschnitt 5.3.2 deren Modifikation für die Ohrmodellausgabe. Einige Merkmale werden mehrmals über kurze Zeitabschnitte, sogenannte *Frames* (Fenster), berechnet. Die verschiedenen Ausprägungen dieser Merkmale für die einzelnen *Frames* werden üblicherweise zusammengefasst. Darauf wird in Abschnitt 5.3.3 genauer eingegangen.

5.3.1 Herkömmliche Merkmale

Im folgenden werden einige der meist verbreitetsten Merkmale für die Instrumentenerkennung definiert (Naghatil und Martin, 2016). Dabei bezeichnet $x[t]$ mit $t = 1, \ldots, M$ eine Beobachtung im Zeitbereich, die aus M Abtastzeitpunkten besteht. Dementsprechend bezeichnet $|X[\mu]|$ die spektrale DFT-Amplitude des μ-ten ($\mu \in \{1, \ldots, M\}$) DFT-Koeffizienten der Beobachtung (siehe Gleichung 2.7). Auf Grund der Symmetrieeigenschaft der DFT ist es ausreichend, lediglich die ersten $M/2$ Fourier-Koeffizienten zu betrachten.

- *Root-Mean-Square (RMS) Energy*: Die globale Energie des Signals $x[t]$ ist die mittlere Energie im Zeitbereich über alle Abtastzeitpunkte $t = 1, \ldots, M$:

$$x_{rms} = \sqrt{\frac{1}{n} \sum_{t=1}^{M} x[t]^2}. \tag{5.1}$$

  Statt der globalen Energie, kann man auch die *Root-Mean-Square Energy* für jeden *Frame* unabhängig messen. Der zeitliche Verlauf dieses Merkmals kann dann beispielsweise zur Schätzung der Länge des Toneinsatzes verwendet werden, z.B durch den Abstand zwischen Tonanfang und dem *Frame* mit maximaler Energie.

- *Lowenergy*: Zunächst wird für jeden *Frame* unabhängig die *Root-Mean-Square Energy* berechnet. Die *Lowenergy* ist dann definiert als der Anteil der *Frames*,

deren Energie kleiner als die mittlere Energie x_{rms} ist. Das Merkmal zeigt auf, wie ungleichmäßig die Energie verteilt ist, beispielsweise nimmt es einen relativ kleinen Wert an, wenn die Amplitude am Anfang des Tons sehr stark ist, anschließend jedoch stark abfällt. Dieses Beispiel zeigt jedoch auch eine Schwäche der *Lowenergy*, denn sie ist auch stark von der Tonlänge abhängig.

- *Spectral Flux*: Dieses Merkmal ist definiert durch die positiven Änderungen des Spektrums benachbarter *Frames* und wurde bereits für die Einsatzzeiterkennung verwendet (vergleiche Gleichung 3.5). Statt der dort verwendeten L1-Norm, wird im Bereich der Instrumentenklassifikation häufiger die L2-Norm verwendet. Das Merkmal ist dann für jeden *Frame* n durch

$$SF(n) = \sum_{\mu=0}^{M/2} H(X[n,\mu] - X[n-1,\mu])^2 \quad \text{mit } H(x) = (x + |x|)/2, \tag{5.2}$$

  definiert. *Spectral Flux* ist ein Maß dafür, wie stark sich die spektrale Zusammensetzung während eines Tons ändert.

- *Zero-Crossing-Rate*: Dieses Merkmal misst, wie oft das Signal das Vorzeichen wechselt, was ein Indikator für einen starken Hochfrequenzanteil oder Rauschen ist. Es ist durch

$$\text{zcr} = \frac{1}{2(M-1)} \sum_{t=1}^{M-1} |sgn(x[t]) - sgn(x(t-1))| \tag{5.3}$$

  definiert, wobei die Vorzeichenfunktion $sgn(\cdot)$ für positive Argumente 1 ergibt und für negative Werte -1. Alternativ kann die *Zero-Crossing-Rate* auch für jeden *Frame* unabhängig berechnet werden.

- *Spectral Centroid*: Dieses Merkmal berechnet den Mittelwert des Spektrums. Es ist definiert durch

$$\text{sc} = \frac{\sum_{\mu=0}^{M/2} \mu \cdot |X[\mu]|}{\sum_{\mu=0}^{M/2} |X[\mu]|}. \tag{5.4}$$

  Da das Merkmal stark von der Tonhöhe abhängt, wird oft auch der *relative Spectral Centroid* verwendet, bei dem der *Spectral Centroid* durch die Tonhöhe (Grundfrequenz) geteilt wird. Hierfür muss aber erst einmal die Tonhöheninformation

vorliegen. Natürlich ist eine Schätzung möglich, führt aber auch zu Fehlern.

- *Spectral Rolloff*: Dieses Merkmal ist definiert als der kleinste Frequenzindex μ_{sr}, unterhalb dessen mindestens ein Anteil von R der kumulierten spektralen Amplitude ($\hat{=}$ die Summe der Amplituden aller Frequenzlinien) konzentriert ist, also das kleinste μ_{sr} für das gilt:

$$\sum_{\mu=0}^{\mu_{sr}} |X[\mu]| \geqslant R \sum_{\mu=0}^{M/2} |X[\mu]|. \tag{5.5}$$

  Ein oft verwendeter Wert für R ist 0.85, wodurch der Hochfrequenzanteil des Signals gemessen wird. Dieser Spezialfall wird als *Spectral Rolloff 85* bezeichnet. Auch der *Spectral Rolloff* ist stark von der Tonhöhe abhängig, weshalb auch hier als Alternative der *relative Spectral Rolloff*, das Verhältnis zwischen *Spectral Rolloff* und Tonhöhe, in Betracht kommt.

- *Spectral Brightness*: Dies ist ein weiteres Merkmal, das den Hochfrequenzanteil eines Signals misst. Es definiert den Anteil der kumulierten spektralen Amplituden der Frequenzen oberhalb einer Frequenz, die üblicherweise auf $f_c = 1500Hz$ gesetzt wird:

$$\mathrm{sb}_c = \frac{\sum_{\mu=\mu_c}^{M/2} |X[\mu]|}{\sum_{\mu=0}^{M/2} |X[\mu]|}, \tag{5.6}$$

  wobei μ_c den Frequenzindex bezeichnet, bei dem die Frequenz f_c enthalten ist. Wie beim *Spectral Centroid* und beim *Spectral Rolloff*, kommt auch hier das Merkmal *relative Spectral Brightness* als Alternative in Betracht, um die Abhängigkeit von der Tonhöhe herauszufiltern.

- *irregularity*: Dieses Merkmal misst die Größe der Variation zwischen den Intensitäten aller benachbarten Partialtöne:

$$irr = \frac{\sum_{n=1}^{N_{\mathrm{part}}} (a_n - a_{n+1})^2}{\sum_{n=1}^{N_{\mathrm{part}}} a_n^2}, \tag{5.7}$$

  wobei N_{part} die Anzahl der betrachteten Partialtöne und a_k die Amplitude des k-ten Partialtons bezeichnet. Je größer der mittlere Unterschied zwischen den benachbarten Partialtönen ist, desto höher wird der Wert von $\mathrm{irr} \in [0, 1]$. Ein bereits mehrfach erwähntes Beispiel für große Unterschiede ist die Klarinette, da bei ihr die Intensitäten fast ausschließlich auf die ungeraden Partialtöne verteilt sind. Für die Berechnung der Intensitäten der Partialtöne ist eine vorherige Grundtonschätzung notwendig, deren Fehler sich natürlich bei der Berechnung der *irregularity*

fortpflanzen.

- Shannon-Entropie: Dieses Merkmal ist durch

$$H(X) = -\sum_{\mu=0}^{M/2} pr(|X[\mu]|) \log_2 pr(|X[\mu]|) \tag{5.8}$$

definiert. Dabei ist $pr(|X[\mu]|) = {}^{|X[\mu]|}/_{\sum_{\nu=1}^{M} |X[\nu]|}$ der Anteil des μ-ten DFT-Koeffizienten bezüglich der kumulierten spektralen Amplitude aller Koeffizienten. $H(X)$ misst den Grad der Zufälligkeit eines akustischen Signals und wird oft als Maß für die Komplexität eines Signals bzw. eines Tons verwendet (Madsen und Widmer, 2006). Die Entropie ist minimal für reine Töne und maximal für weißes Rauschen, bei dem alle Frequenzen identische Amplituden haben.

- Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs beschreiben die spektrale Form eines akustischen Signals bezüglich der menschlichen Wahrnehmung mit Hilfe der Mel-Skala (Davis und Mermelstein, 1980). In dieser Skala werden Tonhöhen so angeordnet, dass deren Distanz der wahrgenommenen Distanz entspricht. Dies entspricht für Frequenzen bis 1000 Hz näherungsweise einer linearen und für höhere Frequenzen einer logarithmischen Abbildung. Die Umrechnung einer Frequenz f (in Hz) in die Mel-Frequenz f_{mel} geschieht gemäß (Stevens, Volkmann und Newman, 1937)

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{5.9}$$

Die gebräuchlichste Definition der MFCC ist auf Basis der DFT unter Verwendung einer Melskalenfilterbank definiert. Zunächst wird das Eingangssignal in *Frames* aufgeteilt, üblicherweise unter Verwendung einer Hamming-Fensterfunktion. Für jeden *Frame* wird dann die DFT berechnet. Anschließend wird der Logarithmus des Amplitudenspektrums berechnet, was damit begründet wird, dass die wahrgenommene Lautstärke in etwa logarithmisch verläuft. Im nächsten Schritt werden die spektralen Komponenten gemäß der Mel-Skala gruppiert und geglättet. Bei der verwendeten Implementierung von Slaney (1998) geschieht dies durch eine Mel-Filterbank mit $N_{\text{mel}} = 40$ halbüberlappenden Bändern. Der Abstand der Mittenfrequenzen der ersten 13 Filter ist äquidistant, während die übrigen 27 Filter logarithmisch angeordnet sind, wobei die Abstände zwischen den Mittenfrequenzen proportional zur Frequenz ist. Die resultierenden Koeffizienten m_j werden logarithmiert und anschließend wird eine Diskrete Cosinus Transformation (DCT)

angewendet:

$$c_{\text{mel}}[i] = \sum_{j=1}^{N_{\text{mel}}} \log(m_j) \cos\left(\frac{\pi i}{N_{\text{mel}}}(j - \frac{1}{2})\right), \text{ für } i = 1, \ldots, N_{\text{mel}}. \tag{5.10}$$

Die DCT dekorreliert die Koeffizienten, wodurch die Merkmalsausprägungen mit diagonalen Kovarianzmatrizen beschrieben werden können (Eronen, 2001). In dem meisten Klassifikationsstudien zur Instrumentenerkennung werden nur die unteren MFCC-Koeffizienten $c_{\text{mel}}[i]$ verwendet, welche die generelle spektrale Struktur erfassen, während die höheren Koeffizienten Informationen über die Tonhöhe und die spektrale Feinstruktur enthalten. Üblicherweise werden die ersten 13 MFCC-Komponenten verwendet (Sandrock, 2013).

5.3.2 Merkmalsextraktion aus der Ohrmodellausgabe

Merkmale auf Basis des Ohrmodells sind eher unüblich und bislang nur in wenigen Studien untersucht. Obwohl solche Merkmale bereits in Martin und Kim (1998) gute Ergebnisse erzielten, sind sie in späteren Studien nicht mehr berücksichtigt worden. Die dort definierten Merkmale beschreiben welche Kanäle wie stark feuern, während die in dieser Arbeit verwendeten Merkmale die spektrale Struktur der Kanalausgaben beschreiben, also in welcher Frequenz die Kanäle feuern. Dabei werden die Kanäle unabhängig berücksichtigt, so dass die Merkmale problemlos auch für *Hearing Dummies*, bei denen bestimmte Kanäle wegfallen (siehe Kapitel 2.1.6), anwendbar sind.

Wie bei der Einsatzzeiterkennung, können auch bei der Instrumentenerkennung die Merkmale, die auf der akustischen Wellenform $x[t]$ definiert sind, auch aus der Ausgabe eines Kanals der Ohrmodellausgabe $p[t, k], k \in \{1, \ldots, K\}$ extrahiert werden. Gleiches gilt entsprechend auch für Merkmale die im Spektralbereich definiert sind, in diesem Fall wird dann $|X[\mu]|$ durch $|P[\mu, k]|, k \in \{1, \ldots, K\}$ ersetzt. Eine Möglichkeit besteht daher auch hier darin, das Musikinstrument von jedem Kanal unabhängig schätzen zu lassen und zum Schluss die Klassifikationsergebnisse zusammenzufassen, z.B. durch eine Majoritätsentscheidung. Da hier jedoch ein klassisches Klassifikationsproblem vorliegt, erscheint eine direkte Aggregation aller Merkmale für die Klassifikationsentscheidung sinnvoller.

In Friedrichs und Weihs (2012) haben wir noch drei weitere Merkmale eingeführt, die auf der Ohrmodellausgabe basieren:

- *One-Cross*: Ein Merkmal, das im auditorischen Kontext wenig Sinn macht, ist *Zero-Cross*, denn hier werden Aktivitätswerte (Spikes/s) betrachtet, die nicht negativ werden können. Statt dessen kann man jedoch zählen, wie oft die Ausgabe den Wert 1 durchschreitet (*One-Cross*), wodurch gemessen wird wie oft die Hörnervenaktivität zwischen schwachen und starken Raten schwankt. Das Merkmal ist dann definiert durch

$$ocr[k] = \frac{1}{2(M-1)} \sum_{t=1}^{M-1} |sgn(p[t,k]-1) - sgn(p(t-1,k)-1)|, \tag{5.11}$$

- *Feuerraten-Mittelwert*: Dieses Merkmal wird über alle Kanäle $k = 1 \dots K$ und alle Zeitpunkte $t = 1 \dots T$ gemittelt:

$$mf = \frac{1}{T \cdot K} \sum_{k=1}^{K} \sum_{t=1}^{T} p[t,k]. \tag{5.12}$$

  Das Merkmal ist jedoch sehr ähnlich zum Merkmal *Root-Mean-Square Energy*. Tatsächlich wäre es identisch, wenn man dort die L2-Norm durch die L1-Norm ersetzen würde und anschließend die Ergebnisse aller Kanäle mittelt.

- *Feuerraten-Varianz*: Dieses Merkmal misst, wie stark sich die Aktivitäten der Kanäle unterscheiden:

$$vf = \mathrm{Var}\left(\frac{1}{T} \sum_{t=1}^{T} p[t,k]\right). \tag{5.13}$$

  Allerdings ist dieses Merkmal auch stark davon abhängig, wie viele Instrumente spielen und es ist daher fraglich, ob es auch für polyphone Musik relevante Information beinhaltet.

5.3.3 Merkmalsaggregation

Von den verwendeten Merkmalen ist lediglich *Spectral Flux* frameweise definiert, das heißt dieses Merkmal liefert eine Vielzahl von Ausprägungen, die aggregiert werden müssen. Für dieses Merkmal werden halbüberlappende *Frames* mit eine Größe von 512 Abtastwerten verwendet, was bei der verwendeten Abtastrate von 44100 Hz 12 ms entspricht. Für eine Tonlänge von beispielsweise 0.2 s entspricht dies einer Anzahl von 33 *Frames* und dementsprechend vielen Merkmalsausprägungen. Diese Werte werden in dieser Arbeit durch den Mittelwert und die Standardabweichung aggregiert. In zukünftigen Untersuchungen

Vorstudie Friedrichs und Weihs (2012)		Alle anderen Experimente	
Nr.	Merkmalsname	Nr.	Merkmalsname
1	Spectral Rolloff 85	1	Spectral Rolloff 85
2	Spectral Brightness	2	Spectral Brightness
3	Irregularity	3	Irregularity
4	Shannon Entropie	4	Shannon Entropie
5 - 17	MFCCs: 13 Koeffizienten	5 - 17	MFCCs: 13 Koeffizienten
18	Feuerraten Mittelwert	18	RMS Energy
19	Feuerraten Varianz	19	Lowenergy
20	One-Cross	20	Spectral Flux Mittelwert
		21	Spectral Flux Varianz

Tabelle 5.1: Verwendete Merkmale für die Instrumentenklassifikation.

könnte dies noch durch eine Aggregation verbessert werden, die den zeitlichen Verlauf des Merkmals berücksichtigt. Beispielsweise unterscheiden sich Gitarren- und Klaviertöne vor allem im Anschlag und nur wenig im späteren Verlauf (Roederer und Mayer, 1999). Daher kann der Unterschied eines Merkmals bei einer reinen Mittelwertbildung über einen langen Ton hinweg sehr gering sein, während der Unterschied im zeitlichen Verlauf eindeutig zu erkennen ist. Zudem sind noch weitere der beschriebenen Merkmale, wie z.B. *RMS Energy* oder *Spectral Rolloff* auch frameweise definierbar und könnten dementsprechend untersucht werden.

Für die Experimente in den Kapiteln 6, 7 und 8 werden, unter Berücksichtigung der Merkmalsaggregation, pro Kanal 21 Merkmale verwendet, die in der rechten Spalte von Tabelle 5.1 aufgelistet sind. Insgesamt ergeben sich somit $41 \cdot 21 = 861$ Merkmale.

5.4 Vorstudien

Vor den Experimenten dieser Arbeit, deren Aufbau in Kapitel 6 beschrieben wird, gab es bereits ein paar experimentelle Vorstudien zu den Ohrmodell-basierten Merkmalen, die in diesem Abschnitt kurz zusammengefasst werden. Für all diese Studien wurde noch eine ältere Version des Ohrmodells von Meddis verwendet, das aus 40 Kanälen mit *Best Frequenzen* zwischen 250 und 7500 Hz besteht (statt 41 Kanäle mit *Best Frequenzen* zwischen 100 und 6000 Hz). In Friedrichs und Weihs (2012) haben wir die kanalbasierten auditorischen Merkmale eingeführt und für monophone Musik getestet (Abschnitt 5.4.1). In einer Bachelorarbeit von Wintersohl (2014) wurden sie auf einem polyphonen Datensatz

gegen die Standardmerkmale (ohne Ohrmodell) getestet (Abschnitt 5.4.2). Für diesen Datensatz werden zudem in Abschnitt 5.4.3 die Merkmalsselektion und in Abschnitt 5.4.4 die hierarchischen Taxonomien getestet.

5.4.1 Friedrichs und Weihs (2012)

In dieser Studie wurden zufällige monophone Tonsequenzen zweier sich zufällig (gleichverteilt) abwechselnder Musikinstrumente verwendet, die mit Hilfe der RWC Datenbank (Goto u. a., 2003) erstellt wurden. Die Klassifikationsaufgabe bestand darin, zu entscheiden, welcher Ton zu welchem der beiden Instrumente gehört. Es wurden drei verschiedene Paare von Instrumenten mit unterschiedlichen Schwierigkeiten getestet: (1) Klavier und Klarinette, (2) Klarinette und Trompete und (3) Klavier und Gitarre. Für jedes Experiment wurden Tonsequenzen mit insgesamt 200 Tönen verwendet. Jeder Ton war 0.5 Sekunden lang, die Tonhöhen und Laustärken wurden zufällig variiert. Wie auch in den Experimenten in Kapitel 6 wurden auch hier die Töne entsprechend ihrer Einsatzzeiten getrennt. Es wurden 20 Merkmale verwendet, die in den linken zwei Spalten von Tabelle 5.1 aufgelistet sind. Neben dem Originalohrmodell wurden auch die Erkennungsraten dreier *Hearing Dummies* (HD) untersucht, die unterschiedlich starke Hörschädigungen nachbilden. HD1 simuliert dabei die stärkste Hörschädigung, während HD2 und HD3 nicht direkt vergleichbar sind (siehe Kapitel 2.1.6). Als Klassifikationsmethode wurde für alle Experimente die lineare SVM verwendet. Die Fehlklassifikationsraten wurden per zehn-mal wiederholter zehnfacher Kreuzvalidierung ermittelt und sind in Tabelle 5.2 zu sehen. Mit dem Modell ohne Hörschädigung wurden akzeptable Fehlerraten zwischen 0,0% und 4,2% erreicht, die jedoch stark von den Instrumentenpaaren abhängen. Die Reihenfolge entsprach dabei der erwarteten Schwierigkeit der Probleme, die durch einen vorherigen informellen Höreindruck eingeschätzt worden waren. Auch bei HD2 und HD3 zeigt sich die gleiche Reihenfolge. Interessanterweise sind die Ergebnisse von HD1, der mit Abstand stärksten Hörschädigung, nicht mehr abhängig von der Klassifikationsaufgabe. Dieser schneidet bei allen drei Experimenten mit Fehlerraten zwischen 31,4% und 32,0% ähnlich schlecht ab. Immerhin sind aber alle Werte immer noch deutlich besser als zufälliges Raten, bei dem eine Fehlerrate von 50% zu erwarten wäre. Weiterhin ist die Reihenfolge der Ergebnisse bezüglich der Hörschädigung konsistent. Für alle drei Klassifikationsaufgaben gilt: $\text{error}_{\text{NH}} < \text{error}_{\text{HD2}} < \text{error}_{\text{HD3}} < \text{error}_{\text{HD1}}$.

	Klavier vs. Klarinette	Klarinette vs. Trompete	Klavier vs. Gitarre
NH	0,0%	0,7%	4,2%
HD 1	31,4%	32,0%	31,5%
HD 2	0,6%	2,4%	6,0%
HD 3	2,0%	6,2%	13,4%

Tabelle 5.2: Fehlklassifikationsraten der Experimente mit den zufälligen Tonsequenzen für das Modell des Normalhörenden (NH) und drei *Hearing Dummies* (HD) (10-mal wiederholte 10-fache Kreuzvalidierung).

5.4.2 Wintersohl (2014)

In dieser Bachelorarbeit wurden die auditorischen Merkmale gegen ihre Standarddefinitionen verglichen. Des Weiteren wurde eine Vielzahl von Klassifikationsmodellen getestet: Entscheidungsbäume (CART), lineare Diskriminanzanalyse (LDA), quadratische Diskriminanzanalyse (QDA), lineare SVM (SVML), radiale SVM (SVMR), polynomielle SVM (SVMP), *Random Forest* (RF) und k-Nächste-Nachbarn (k-NN). Neben monophonen Einzeltönen wurde auch die predominante Instrumentenerkennung anhand von Ausschnitten aus willkürlich gewählten Kammermusikstücken getestet. Jedes Stück wurde in fünf verschiedenen Variationen mit einem jeweils anderen Melodieinstrument (Flöte, Klarinette, Oboe, Trompete und Violine) aus einer MIDI-Version in das Wave-Format synthetisiert. Insgesamt bestehen die Musikausschnitte aus 170 Melodietönen, so dass sich insgesamt $5 \cdot 170 = 850$ Beobachtungen ergeben. Im Gegensatz zu dem Versuchsaufbau in Kapitel 6 ist bei diesen Daten nicht sichergestellt, dass das Melodieinstrument auch tatsächlich dominant ist.

Merkmale	CART	LDA	QDA	SVMR	SVML	SVMP	RF	k-NN
Standard	0.36	0.25	0.24	0.21	0.21	**0.19**	0.22	0.27
Ohrmodell	0.32	0.78	-	0.15	0.15	**0.14**	0.16	0.27

Tabelle 5.3: Fehlklassifikationsraten predominanter Instrumentenerkennung für 5 Melodieinstrumente (10-fache Kreuzvalidierung) (Wintersohl, 2014).

Die Ergebnisse der Bachelorarbeit sind in Tabelle 5.3 zu sehen. Es zeigt sich, dass die auditorischen Merkmale für fast alle Klassifikationsverfahren besser abschneiden als die Standardmerkmale. Ausnahmen sind lediglich die Methoden LDA und QDA, die ein Problem mit kollinearen Merkmalen bei der Ohrmodellvariante haben. Dieses Problem könnte durch eine vorgeschaltete Variablenselektion behoben werden. Am besten schneidet die polynomielle SVM ab, die eine Fehlklassifikationsrate von 14%

erreicht. Nur minimal schlechter ist allerdings die lineare SVM, deren Laufzeit für die Hyperparameteroptimierung deutlich kürzer ist, da nur ein Parameter angepasst werden muss.

5.4.3 Experiment zur Merkmalsselektion

Durch die gruppenbasierte Variablenselektion, die in Kapitel 2.4 beschrieben ist, lassen sich die Ergebnisse aus Wintersohl (2014) für die Ohrmodellmerkmale noch weiter verbessern. Die Ergebnisse sind in Tabelle 5.4 aufgelistet. Besonders gut schneidet die Rückwärtsselektion unter Verwendung der Kanalgruppierung (alle Merkmale des gleichen Kanals gehören zu einer Gruppe) ab. In diesem Fall erreicht die polynomielle SVM eine Fehlerrate von 12% (statt 14%).

Selektionsmethode	CART	LDA	QDA	SVMR	SVML	SVMP	RF	k-NN
keine Selektion	0.32	0.78	-	0.15	0.15	0.14	0.16	0.27
normale VS	0.30	0.21	0.22	0.21	0.20	0.22	0.20	0.21
Kanal-basierte VS	0.29	0.17	0.23	0.16	0.18	0.17	0.17	0.20
Kanal-basierte RS	0.28	0.75	-	0.13	0.14	**0.12**	0.15	0.18
Merkmalstyp-basierte VS	0.32	0.18	0.28	0.15	0.18	0.16	0.17	0.20
Merkmalstyp-basierte RS	0.30	0.18	-	0.14	0.15	0.14	0.15	0.21

Tabelle 5.4: Fehlerraten der Merkmalsselektion für die Ohrmodellmerkmale und den Datensatz aus Wintersohl (2014) (VS = Vorwärtsselektion, RS = Rückwärtsselektion).

5.4.4 Experiment zum Taxonomiedesign

Hier wird der Datensatz aus Wintersohl (2014) für den Test der verschiedenen Taxonomien verwendet, die in Abschnitt 5.2 vorgestellt sind. Neben einer flachen Taxonomie ist dies die hierarchische Taxonomie nach Instrumentenfamilie, sowie die Hornbostel-Sachs-Taxonomie. Für die fünf Instrumente des Datensatzes – Flöte, Klarinette, Oboe, Trompete und Violine – sind die beiden hierarchischen Varianten in den Abbildungen 5.5 und 5.6 visualisiert.

Für jeden Knoten wird das individuell beste Klassifikationsmodell gewählt (entweder lineare SVM oder *Random Forest*). Die Ergebnisse in Tabelle 5.5 zeigen jedoch, dass die hierarchischen Taxonomien weder für die Ohrmodellmerkmale noch für die Standardmerkmale eine Verbesserung im Vergleich zur flachen Taxonomie bringen. Möglich ist allerdings, dass eine individuelle Merkmalsselektion an jedem Knoten die Ergebnisse

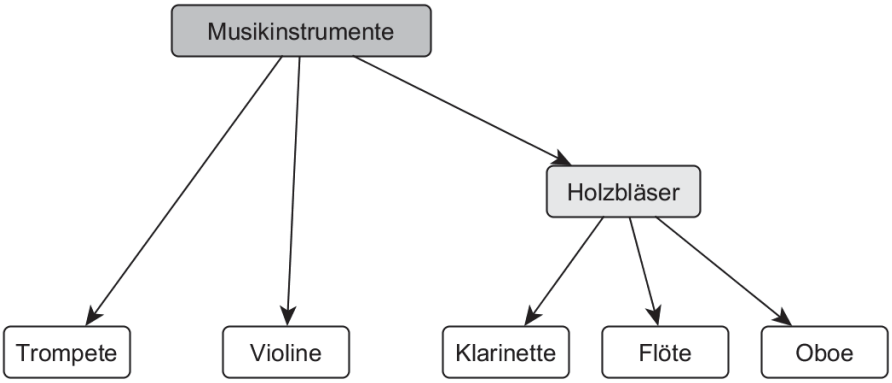

Abbildung 5.5: Hierarchische Taxonomie nach Instrumentenfamilie für das untersuchte Beispiel.

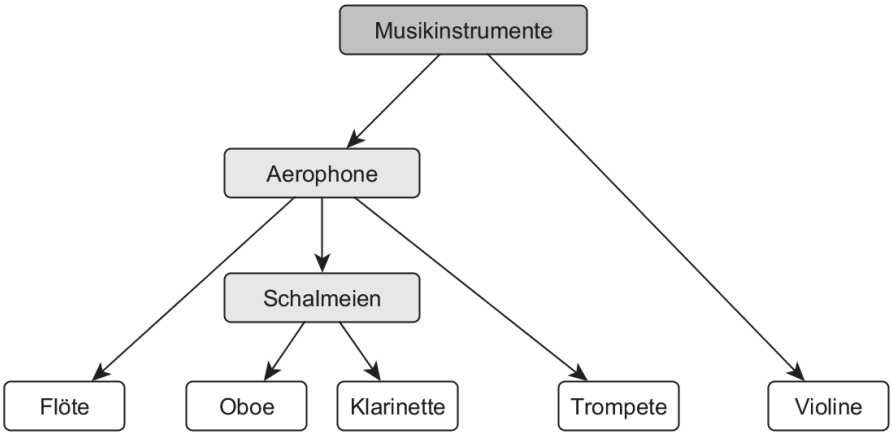

Abbildung 5.6: Hornbostel-Sachs-Taxonomie für das untersuchte Beispiel.

der hierarchischen Klassifikation verbessern könnte. Die individuellen Fehlerraten der einzelnen Knoten sind in den Tabellen A.3 und A.4 im Anhang A aufgelistet.

Merkmale	Taxonomie	bestes kombinierte Modell
Standard	flache Taxonomie	0.21
	Taxonomie nach Instrumentenfamilien	0.23
	Hornbostel-Sachs-Taxonomie	0.21
Ohrmodell	flache Taxonomie	**0.15**
	Taxonomie nach Instrumentenfamilien	**0.15**
	Hornbostel-Sachs-Taxonomie	0.16

Tabelle 5.5: Gesamtfehlerraten der drei Taxonomien.

6 Versuchsdesign

Die in den Kapiteln 3, 4 und 5 beschriebenen Verfahren zur Einsatzzeit-, Tonhöhen- und Instrumentenerkennung sollen in einem gemeinsamen Versuch getestet werden. In diesem Kapitel wird dafür zunächst in Abschnitt 6.1 ein Versuchsplan erstellt. Dieser Plan definiert, welche Daten für die Experimente verwendet werden. Anschließend wird in Abschnitt 6.2 erläutert, welche Vergleichsexperimente durchgeführt und welche Verfahren berücksichtigt werden. Am Ende dieses Kapitels wird in Abschnitt 6.3 die verwendete Software beschrieben.

6.1 Versuchsplan zur Datenauswahl

Der Datenauswahl liegt eine Datenbank zugrunde, die aus 100 Kammermusikstücken im MIDI-Format besteht. Jedes dieser Stücke besteht aus einer Melodiestimme und einer oder mehrerer Begleitstimmen. Das Melodieinstrument ist entweder Cello, Klarinette oder Trompete, und die Begleitung ist entweder Klavier oder Streicher. Die ISP Toolbox (Jensen, Christensen und Jensen, 2007) in Matlab wird verwendet um die MIDI-Stücke in Wave-Dateien mit einer Abtastrate von 44100 Hz umzuwandeln. Reale Musikaufnahmen wären natürlich wünschenswerter, aber das gewählte Konzept bietet eine Datenbasis gelabelter Musikdaten – mit Einsatzzeiten, Tonhöhen und Musikinstrumenten – die ausreichend groß für den Versuchsplan ist. Bei MIDI-Stücken ist die Information über die beteiligten Musikinstrumente, Tonanfangszeiten, Tondauern, Tonhöhen und Lautstärken direkt ablesbar und zudem auch einfach modifizierbar.

In den meisten Studien der Musikdatenanalyse werden die Musikdaten für die Experimente relativ willkürlich ausgewählt, so dass es schwierig ist zu bestimmen, wie gut diese Daten die Gesamtheit der Musik repräsentieren. In dieser Arbeit werden stattdessen die Musikdaten mit Hilfe eines Versuchsplans in einer strukturierteren Form ausgewählt. Für die Generierung dieses Versuchsplans werden acht Faktoren definiert, durch die ein Musikstück (oder Ausschnitt) in einem achtdimensionalen Raum lokalisiert werden kann.

Im Anschluss an die Experimente kann ein Regressionsmodell angepasst werden, das einen Zusammenhang zwischen den Faktoren und den Fehlklassifikationskosten herstellt. So kann beispielsweise ermittelt werden, ob es einen bestimmten Musiktyp gibt, für den ein Verfahren besonders gute oder schlechte Ergebnisse liefert. Die Faktoren sind so ausgewählt, dass sie möglichst gut die wesentlichen Eigenschaften der Musik abdecken. Trotz allem können Musikstücke natürlich nicht vollständig durch acht Dimensionen beschrieben werden. Für die Regressionsmodelle ist daher mit relativ großen Störgrößen zu rechnen, die die nicht berücksichtigten Faktoren enthalten.

Die Faktoren können in zwei Gruppen aufgeteilt werden: Faktoren, deren Veränderung ein Musikstück so verändert, dass es anschließend unnatürlich klingt und Faktoren, deren Änderung die musikalische Struktur weitestgehend beibehält. Da hier nur reale Musik von Belang sein soll, ist die erste Faktorgruppe kritisch. Die Werte dieser Faktoren gelten daher im Folgenden als unveränderbar. Dagegen werden die Faktoren der zweiten Gruppe entsprechend der Vorgabe des Versuchsplans angepasst. Von den hier berücksichtigten acht Faktoren sind vier der ersten Gruppe zuzurechnen und die restlichen vier der zweiten Gruppe.

Da die Vorverarbeitung, bestehend aus der Transformation des Ohrmodells und der Merkmalsextraktion, sehr rechenintensiv ist, muss die Menge der Daten möglichst klein gewählt werden. Andererseits muss sie aber hinreichend groß für die Anpassung des Regressionsmodells sein. Zudem ist auch die Güte der Klassifikationsmodelle von der Datengröße abhängig. Einen guten Kompromiss bilden Plackett-Burman (PB) Versuchspläne, die mit relativ wenigen Versuchen ($\widehat{=}$ Musikstücken) auskommen. Für diese Pläne müssen für jeden Faktor zwei Niveaus definiert werden, ein hohes (+) und ein tiefes (−). Statt kompletter Musikstücke, werden aus Rechenzeitgründen nur kurze Ausschnitte verwendet. Diese sind so definiert, dass sie den Zeitbereich von 30 Tönen der Melodie abdecken. Ein Nachteil der PB-Versuchspläne ist allerdings, dass sie keine Wechselwirkungen zwischen den Faktoren aufdecken können. Diese sind jedoch für die acht Faktoren nicht vollkommen auszuschließen.

Die Faktoren der ersten Gruppe sind im Folgenden definiert. Da deren Ausprägungen unveränderbar sind, werden Intervalle für die Niveaus des PB-Designs definiert. Diese werden so gewählt, dass beide Niveaus eine in etwa gleich große Menge von Musikstücken definieren, und zudem auch ein hinreichend großer Abstand zwischen den Intervallen besteht.

- **Mittleres Intervall**: Dieser Faktor ist definiert durch den mittleren Tonhöhenabstand, in Halbtönen gemessen, zwischen zwei benachbarten Tönen der Melodiestimme. Es werden zwei Niveaus definiert: < 2.5 und > 3.5.

- **Anteil der zusätzlichen Toneinsätze durch die Begleitung**: Dieser Faktor definiert den Anteil der individuellen Einsatzzeiten der Begleitinstrumente, die nicht in der Melodiestimme vorkommen, im Verhältnis zu allen Toneinsätzen. Die zwei verwendeten Niveaus sind definiert durch < 0.4 und > 0.6.

- **Dynamik**: Die Dynamik eines Musikstücks wird hier definiert durch die mittleren Lautstärkeunterschiede zweier benachbarter Töne, gemessen in MIDI-Laustärkenummern. Die beiden verwendeten Niveaus sind definiert durch < 0.5 und > 1.0.

- **Begleitinstrument**: Es kommen zwei verschiedene Begleitinstrumente (Niveaus) vor: Klavier und Streicher.

Die vier Faktoren der zweiten Gruppe, die innerhalb eines Definitionsbereichs frei einstellbar sind, sind wie folgt definiert:

- **Melodieinstrument**: Es werden drei verschiedene Musikinstrumente untersucht, die aus unterschiedlichen Instrumentenfamilien stammen: Cello, Klarinette und Trompete. Da hier 3 Faktoren vorkommen, wird dieses Merkmal außerhalb des Plackett-Burman-Plans berücksichtigt. Stattdessen wird jedes Experiment (Musikstück) des Plans dreimal wiederholt, einmal für jedes Instrument.

- **Mittlere Tonhöhe der Melodie**: Die minimal und maximal erlaubten Tonhöhen sind angepasst an den gemeinsamen Tonhöhenumfang der drei verwendeten Musikinstrumente: von E3 (165 Hz) bis A6 (1047 Hz). Die zwei Niveaus der mittleren Tonhöhe sind definiert durch D4 (294 Hz) und D5 (587 Hz). Gemäß dieser werden die Musikstücke entsprechend transponiert. Hierbei kann jedoch für einzelne Töne die Beschränkung durch die minimale oder maximale Tonhöhe verletzt werden. In diesem Fall wird das komplette Stück so weit transponiert, wie es nötig ist, diese Verletzung zu beheben.

- **Tonlänge**: Aus technischen Gründen wird die Tonlänge durch die Gesamtlänge des Musikstücks definiert. Die Länge der Musikausschnitte wird entsprechend der beiden Niveaus, 12 s und 25 s, gestreckt oder komprimiert. Für die Musikausschnitte der verwendeten Datenbank ergibt dies dann Tonlängen im Bereich von $[0.1, 0.5]$ s für das erste Niveau und von $[0.2, 1.0]$ s für das zweite Niveau.

- **Mittlere Tonhöhe der Begleitung (im Verhältnis zur Melodie)**: Dieser Faktor ist definiert als die Differenz von der mittleren Tonhöhe der Begleitung zu der mittleren Tonhöhe der Melodie. Damit der Klang des Musikstücks nicht zu stark verändert wird, wird hierbei nur eine Transposition der Begleitstimmen um vollständige Oktaven (zwölf Halbtöne) erlaubt. Die Niveaus sind daher auf Intervalle der Länge zwölf definiert: [-6,6] und [-24,-12]. Wenn die Tonhöhen von Melodie und Begleitung ähnlich sind, überdecken sich mehr Obertöne und die Trennung der beiden Stimmen wird schwieriger. Der Fall, in dem die Begleitung deutlich höher als die Melodie ist, wird ignoriert, da dies in realen Musikstücken relativ selten vorkommt.

Faktor	1. Niveau (−)	2. Niveau (+)
Mittleres Intervall	<2.5	>3.5
Toneinsätze der Begleitung	<0.4	>0.6
Dynamik	<0.5	>1.0
Begleitinstrument	Klavier	Streicher
Mittlere Tonhöhe	D4	D5
Tonlänge (Länge des Musikstücks)	12 s	25 s
Tonhöhendifferenz: Melodie - Begleitung.	$[-6, 6]$ Halbtöne	$[12, 24]$ Halbtöne

Tabelle 6.1: Plackett-Burman-Design: Niveauwerte der Faktoren.

Die Faktoren des PB-Designs und ihre definierten Niveaus sind in Tabelle 6.1 zusammengefasst. Es werden PB-Designs mit zwölf Versuchen verwendet, wobei jeder Versuch eine bestimmte Kombination der Faktorniveaus definiert. Für jeden Versuch werden aus der Gesamtdatenbank alle Musikausschnitte mit einer Länge von 30 Melodietönen ausgewählt, die den Niveaus der ersten Faktorgruppe entsprechen. Anschließend wird eines dieser Stücke ausgewählt und die Faktoren der zweiten Gruppe entsprechend angepasst. Zum Schluss werden drei Versionen dieses Musikausschnitts erstellt, jede mit einem anderen der drei Musikinstrumente für die Melodie. Insgesamt ergeben sich somit $3 \cdot 12 \cdot 30 = 1080$ Melodietöne für ein Plackett-Burman-Design. Es werden drei verschiedene Designs erstellt – jedes bestehend aus anderen Musikausschnitten –, wodurch eine Evaluation mittels einer dreifachen Kreuzvalidierung ermöglicht wird. Um sicherzustellen, dass die Melodie auch in allen Stücken ähnlich dominant ist, wird ein Melodie-/Begleit-Verhältnis von 5 dB verwendet. Für die Auswertung des Versuchsplans ist zu beachten, dass die Einsatzzeiterkennung ein Maximierungsproblem ist (F-Maß), während Tonhöhen- und Instrumentenerkennung Minimierungsprobleme sind (Fehlerrate). Dementsprechend werden in den beiden Fällen gleiche Faktoreffekte mit unterschiedlichen Vorzeichen angezeigt.

6.2 Aufbau der Vergleichsexperimente

Zunächst werden die verschiedenen Verfahren für das Ohrmodell ohne Hörschädigung getestet und mit anerkannten Standardverfahren (ohne Ohrmodell) verglichen. Die Struktur dieses Prozesses ist zusammenfassend in Abbildung 6.1 dargestellt.

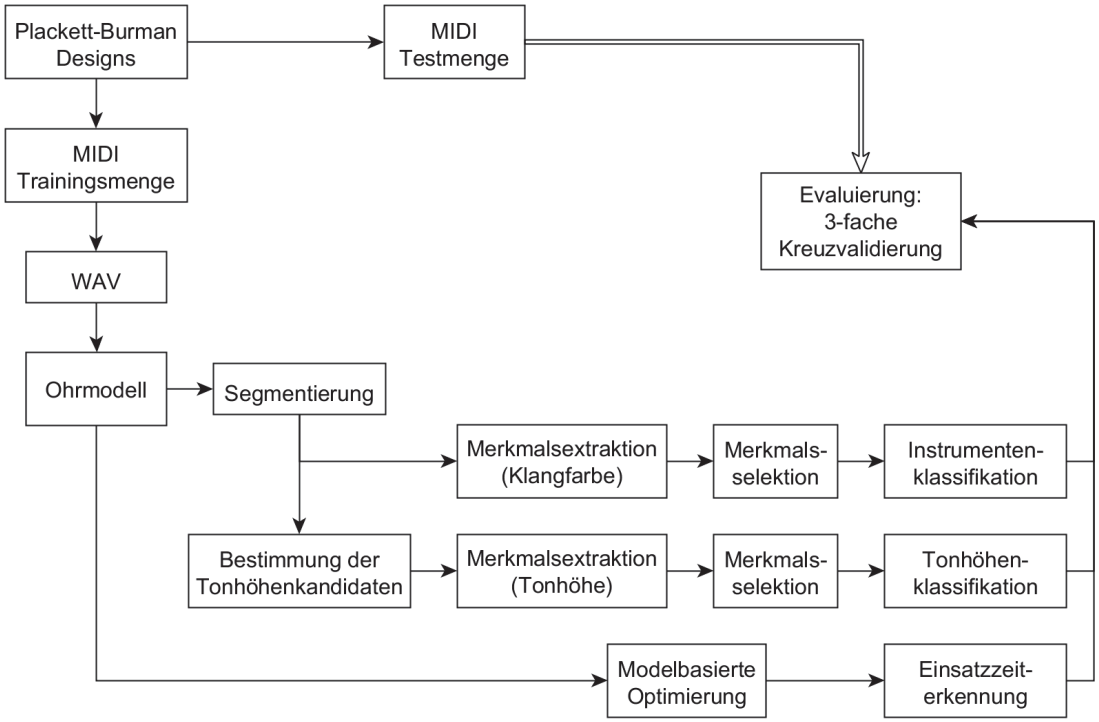

Abbildung 6.1: Aufbau der Experimente für die Verfahren zur Musikerkennung mit Ohrmodell.

Für die Evaluation der Fehlerraten wird dreifache Kreuzvalidierung verwendet, wobei in jeder Iteration zwei PB-Designs für die Trainingsmenge und ein PB-Design für die Testmenge verwendet werden. Die MIDI-Daten der Testmenge werden zu Wave-Daten synthetisiert und anschließend durch das Ohrmodell in multivariate Spikefeuerraten transformiert. Auf Basis dieser Feuerraten werden Klassifikationsmodelle trainiert (bei der Einsatzzeiterkennung entspricht dies der Optimierung). Schließlich werden die Modelle durch die Testdaten, die vorher wie die Trainingsdaten transformiert werden, evaluiert.

Von Interesse sind vor allem die predominanten Erkennungsraten der verschiedenen Verfahren. Zusätzlich werden aus Vergleichszwecken auch alle Experimente mit monophonen

Daten getestet. Dabei werden die gleichen Daten leicht modifiziert verwendet, indem die Begleitstimmen aus den MIDI-Stücken entfernt werden. Ohne die Verzerrung durch die Begleitung, sollten die Fehlerraten für diese Musikdaten minimal sein.

Da die predominante Variante der Einsatzzeiterkennung bisher noch nicht näher untersucht worden ist, ist es schwierig die Leistung der hierfür vorgeschlagenen Verfahren einzuordnen. Die Standardvarianten sind dagegen die monophone Variante oder die polyphone Variante, in der die Einsatzzeiten aller Stimmen gesucht werden. Daher wird neben der monophonen und der predominanten Variante auch die übliche polyphone Variante getestet, um so die Einsatzzeiterkennung mit Ohrmodell besser bewerten zu können. Alle 12 Kombinationen der Einsatzzeiterkennung – drei Verfahren mit und eins ohne Ohrmodell, kombiniert mit den 3 Varianten – werden individuell mit sequentieller Modellbasierter Optimierung (MBO) optimiert, wobei jeweils ein Budget von 275 Iterationen vorgegeben ist, das heißt 275 verschiedene Parameterkombinationen werden in der Trainingsphase getestet.

Für Tonhöhen- und Instrumentenerkennung werden alle Verfahren mit den zwei in Kapitel 2.3 beschriebenen Klassifikationsverfahren getestet: *Random Forest* und lineare SVM. Für die Instrumentenerkennung bedeutet dies einen Vergleich von vier Varianten: Die Ohrmodellmerkmale werden gegen die herkömmlichen Merkmale getestet (siehe Kapitel 5.3), und beide Varianten werden mit den zwei Klassifikationsmethoden verknüpft. Für die Tonhöhenschätzung werden insgesamt sieben Verfahren getestet: Vier Varianten mit Klassifikation – *Random Forest* oder lineare SVM, und DFT- oder ACF-Merkmale (siehe Kapitel 4.2.2) –, zwei Varianten zur Peakauswahl für den SACF-Ansatz (siehe Kapitel 4.1.2) und der YIN-Algorithmus (siehe Kapitel 4.1.1), das Standardverfahren ohne Ohrmodell. Allerdings wird dieses Verfahren lediglich mit den voroptimierten Parametereinstellungen verwendet und nicht für die spezielle Struktur der Musikdaten optimiert. Dementsprechend sind die Ergebnisse dieses Verfahrens etwas beeinträchtigt.

Zudem werden für Tonhöhen- und Instrumentenerkennung die beiden in Kapitel 2.4 beschriebenen Verfahren zur Merkmalsselektion angewendet, um die Wichtigkeit von Kanälen und Merkmalen zu ermitteln. Alle beschriebenen Experimente für die Verfahren mit Ohrmodell werden auch für die drei in Kapitel 2.1.6 beschriebenen *Hearing Dummies* durchgeführt. Hier soll untersucht werden, ob die gemessenen Hörbeeinträchtigungen bezüglich der drei Musikwahrnehmungsaufgaben plausibel erscheinen.

Schließlich werden für alle Experimente die Versuchspläne ausgewertet. Dabei wird analysiert, ob es bestimmte Eigenschaften der Musik gibt, durch die die Erkennung

einfacher oder schwieriger wird. Diese Erkenntnisse können auch auf Unterschiede zwischen verschiedenen Verfahren oder Hörschädigungen untersucht werden. Auf Grund des hohen Rechenzeitbedarfs für die Ohrmodelltransformation, die Merkmalsextraktion und die Optimierung der Einsatzzeiterkennung, werden alle Experimente nur einmal ausgeführt.[1]

6.3 Verwendete Software

Die **ISP Toolbox** in Matlab wird genutzt, um MIDI-Daten in Wave-Daten mit einer Abtastrate von 44100 Hz umzuwandeln (Jensen, Christensen und Jensen, 2007). Für die Ohrmodellsimulation wird das **Ohrmodell von Meddis** in Matlab mit Standardeinstellungen (41 Kanäle, *Best Frequenzen* von 100 Hz bis 6 kHz) angewendet (Meddis, 2006; Panda u. a., 2014). Für den **YIN-Algorithmus** wird die Matlab-Implementierung der Autoren mit Standardeinstellungen verwendet (De Cheveigné und Kawahara, 2002). Bis auf die selbst programmierte Shannon-Entropie, werden alle anderen 20 Merkmale für die Instrumentenklassifikation mit Hilfe der **MIRtoolbox** in Matlab extrahiert (Lartillot und Toiviainen, 2007). Die Klassifikationsaufgaben werden mit dem R-Paket **mlr** (Bischl u. a., 2016a) durchgeführt, wobei die Pakete **randomForest** (Liaw und Wiener, 2002) für *Random Forest* und **kernlab** (Karatzoglou u. a., 2004) für die linear SVM verwendet werden. Für den *Random Forest* werden die Standardeinstellungen genutzt ($ntree = 500$), das heißt für jedes Modell werden 500 Bäume verwendet. Für die lineare SVM wird der in einer Vorstudie optimierte Parameterwert $cost = 1$ verwendet, wobei *cost* dem Parameter C in Kapitel 2.3.2 entspricht. Für die sequentielle modellbasierte Optimierung wird das R-Pakets **mlrMBO** (Bischl u. a., 2016b) mit folgenden Einstellungen genutzt:

- Größe des Startdesigns: $5p$ (55 für den 1. MBO-Lauf, 20 für den 2. MBO-Lauf),
- Anzahl an Iterationen: $20p$ (220 für den 1. MBO-Lauf, 80 für den 2. MBO-Lauf),
- Surrogatmodell: Kriging mit Matern 3/2 Kernel,
- Infill-Kriterium: *Expected Improvement* (EI),
- Infill-Optimierer: *Focus Search*.

Die Einsatzzeiterkennung basiert auf einem R-Code, der in Bauer (2016) beschrieben ist, und der für die Spektralanalyse das R-Paket **tuneR** (Ligges u. a., 2014) verwendet. Die

[1] Trotz einer parallelen Ausführung, dauerten die Experimente über einen Monat.

hohe Anzahl an Experimenten wird mit den R-Paketen **BatchJobs** und **BatchExperiments** organisiert (Bischl u. a., 2015).

7 Ergebnisse der Vergleichsexperimente

Zunächst werden die Fehlerraten der entwickelten Verfahren präsentiert und mit den Ergebnissen der Standardverfahren verglichen (Abschnitt 7.1). Dafür wird jeweils das Ohrmodell ohne Hörschädigung verwendet. Im zweiten Teil des Kapitels werden die Leistungsverluste untersucht, die durch die Verwendung der Modelle mit Hörschädigung entstehen (Abschnitt 7.2).

7.1 Vergleich der Ohrmodellverfahren zu Standardverfahren

Im Folgenden werden die Ergebnisse der Verfahren für die Einsatzzeiterkennung, die Tonhöhenschätzung und die Instrumentenerkennung erläutert.

7.1.1 Einsatzzeiterkennung

Tabelle 7.1 zeigt die Ergebnisse der Einsatzzeiterkennung bezüglich der vier untersuchten Methoden: (1) Standardverfahren der Einsatzzeiterkennung (ohne Ohrmodell), (2) beste Einkanalschätzung, (3) quantilbasierte Aggregation der Merkmalsvektoren und (4) Aggregation der Einkanalschätzungen. Für alle Methoden sind die relevanten Parameter separat bezüglich der drei untersuchten Varianten – monophone, predominante und polyphone Einsatzzeiterkennung – optimiert.

Alle Verfahren schneiden in allen Varianten eher schlecht ab. Auch das Standardverfahren erreicht in der monophonen Variante lediglich einen mittleren F-Wert von 0.86. Demgegenüber dürfte ein Normalhörender bei diesen Musikdaten, bei denen alle Töne eine Mindestlänge von 0.1 s haben, relativ problemlos alle Toneinsätze fehlerfrei identifizieren können. Ein F-Wert von 0.86 bedeutet indes beispielsweise gemäß Gleichung 3.10, dass 24.6% aller Toneinsätze nicht erkannt werden, wenn es keinen Fehlalarm gibt ($FP = 0$).

Verfahren	Instrument	Polyphon	Predominant	Monophon
Standardverfahren (ohne Ohrmodell)	Cello	0.65	0.57	0.80
	Klarinette	0.79	0.72	0.80
	Trompete	0.87	0.84	0.97
	Mittelwert	**0.77**	**0.71**	**0.86**
beste Einkanalschätzung	Cello	0.44	0.37	0.68
	Klarinette	0.65	0.61	0.80
	Trompete	0.70	0.79	0.99
	Mittelwert	**0.60**	**0.59**	**0.82**
quantilbasierte Aggregation	Cello	0.40	0.34	0.60
	Klarinette	0.57	0.59	0.81
	Trompete	0.72	0.76	0.99
	Mittelwert	**0.56**	**0.56**	**0.80**
Aggregation der Einkanalschätzungen	Cello	0.53	0.46	0.79
	Klarinette	0.71	0.72	0.76
	Trompete	0.85	0.87	0.98
	Mittelwert	**0.69**	**0.68**	**0.84**

Tabelle 7.1: Ergebnisse(mittlerer F-Wert) für die Einsatzzeiterkennung mit und ohne Ohrmodell.

Bei den Einzelkanalschätzungen sind die Ergebnisse sehr unterschiedlich, je nachdem welcher Kanal verwendet wird, wie in Abbildung 7.1 zu sehen ist. Für die predominante Aufgabe sind die mittleren Kanäle besser als Kanäle mit niedrigen oder hohen Mittenfrequenzen. Die kleinste Fehlerrate erzielt der Kanal 14 mit einem mittleren F-Wert von 0.59. Für die monophone Einsatzzeiterkennung schneidet Kanal 15 mit einem mittleren F-Wert von 0.82 am besten ab. In dieser Variante erzielen die hohen Kanäle ab Nummer 25 deutlich schlechtere Ergebnisse. Sehr ähnlich sieht das Ergebnis auch für die polyphone Einsatzzeiterkennung aus, nur mit deutlich höheren Fehlerwerten. Hier schneiden die Kanäle 10 und 17 mit mittleren F-Werten von 0.60 am besten ab.

Die quantilbasierte Aggregation verschlechtert die Ergebnisse für alle Varianten. Problem dieses Verfahrens sind die unterschiedlichen Latenzen der Kanäle (vergleiche Abbildung 3.2 in Kapitel 3.3), die in dem Verfahren nicht kompensiert werden. Für die Zukunft sollte dieses Verfahren daher um einen entsprechenden Mechanismus erweitert werden.

Die Aggregation der Einkanalschätzungen verbessert dagegen die Ergebnisse deutlich. Bei diesem Verfahren werden die unterschiedlichen Latenzen durch den Verschiebungsparameter τ berücksichtigt, der für jeden Kanal unabhängig optimiert wird. Für die predominante Einsatzzeiterkennung werden im Optimum überraschend alle Kanäle be-

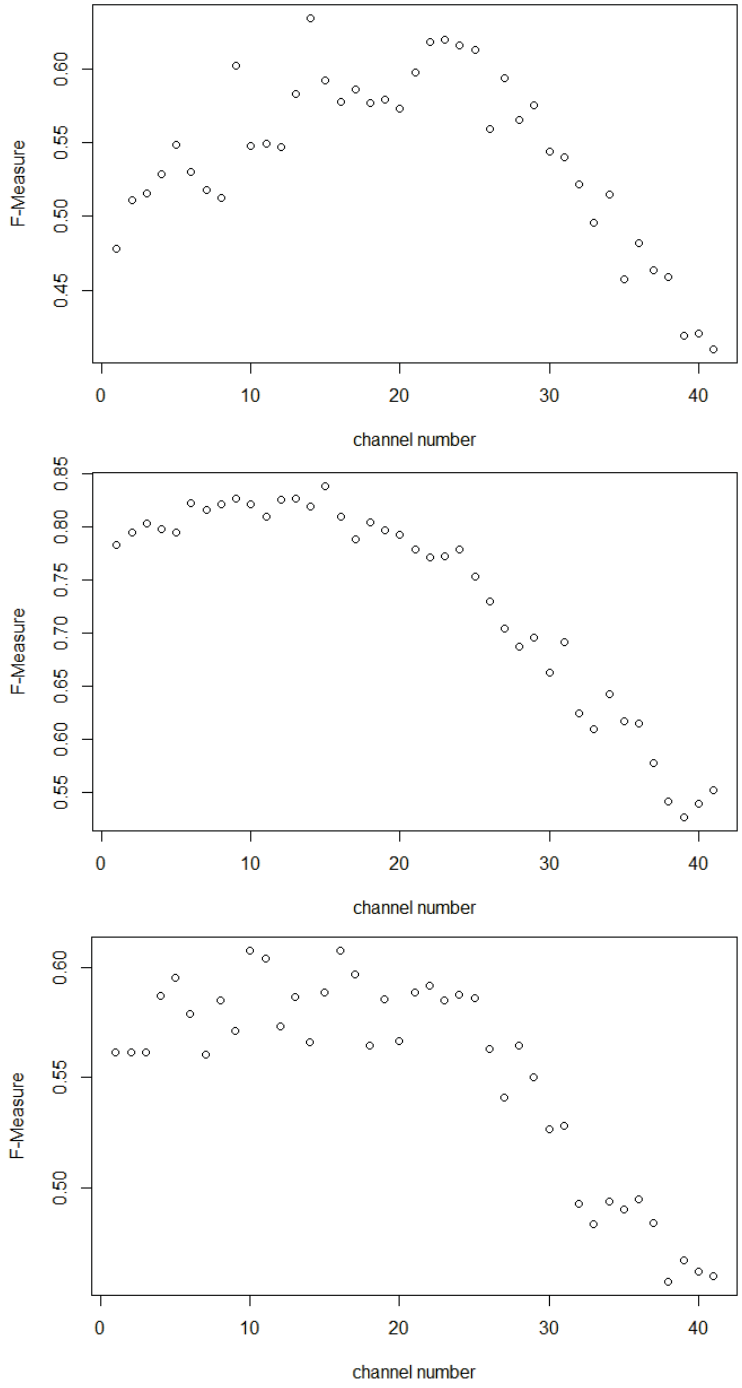

Abbildung 7.1: Ergebnisse (mittlerer F-Wert) für die Einsatzzeiterkennung unter Verwendung lediglich eines Kanals – Oben: predominant, Mitte: monophon, Unten: polyphon.

rücksichtigt ($k_{\min} = 1$ und $k_{\max} = 41$), obwohl die niedrigen und hohen Kanäle bei der Einkanalschätzung nicht gut abschneiden (vergleiche Abbildung 7.1). Der mittlere F-Wert von 0.68 ist allerdings immer noch etwas schlechter als das Standardverfahren ohne Ohrmodell, das einen mittleren F-Wert von 0.71 erreicht.

Die Trompete schneidet in allen getesteten Varianten und für alle Methoden mit Abstand am besten ab, wohingegen Einsatzzeiten des Cellos am schlechtesten erkannt werden. In der predominanten Variante ist deren Erkennung bei einer Streicherbegleitung zusätzlich erschwert, da in diesem Fall sehr ähnliche Einsätze differenziert betrachtet werden müssen. Bei einem Vergleich der Fehlerraten verschiedener Instrumente muss beachtet werden, dass immer nur die Gesamtleistung eines Verfahrens bezüglich aller Instrumente optimiert ist. Dabei wird manchmal eine etwas schlechtere Leistung für ein Instrument in Kauf genommen, wenn dadurch die Erkennungsraten für ein anderes Instrument erhöht wird. Demzufolge kann man beispielsweise nicht folgern, dass Trompeteneinsätze von den Ohrmodellverfahren besser erkannt werden als vom Standardverfahren, obwohl die F-Werte darauf hindeuten (0.99 gegenüber 0.97).

Wie zu erwarten sind bei der polyphonen und der predominanten Einsatzzeiterkennung die Ergebnisse aller Verfahren schlechter als bei der monophonen Variante. Bemerkenswert ist allerdings, dass die Suche nach allen Einsatzzeiten einfacher ist als die ausschließliche Suche nach den Melodietoneinsätzen. Demzufolge stellt die Trennung zwischen Melodie- und Begleiteinsätzen ein schwerwiegendes Problem dar. Dies gilt zumindest für das verwendete Melodie-/ Begleitverhältnis von 5 dB.

Tabelle 7.2 zeigt die Auswertung des Versuchsplans für das beste Verfahren der Einsatzzeiterkennung, die Aggregation der Einzelschätzungen, gemittelt über alle Designs und alle Instrumente. In der monophonen Variante ist das adjustierte R-Quadrat negativ, was darauf hinweist, dass der Fehler (genauer der F-Wert) unabhängig von der Art der Musik ist. Dies wird auch durch die p-Werte bestätigt, denn keiner zeigt einen signifikanten Einfluss an. Für die Faktoren, die nur im Zusammenhang mit der Begleitung eine Bedeutung haben, ist dies natürlich offensichtlich. Aber bemerkenswert ist, dass auch größere Intervalle und Lautstärkeunterschiede die Einsatzzeiterkennung nicht signifikant erleichtern.

Für die beiden anderen Varianten der Einsatzzeiterkennung ist die Anpassungsgüte der Auswertung relativ hoch ($R_a^2 > 0.5$).[1] Hier können auch Einflussfaktoren ermittelt

[1] Bei den Anpassungsgüten ist zu beachten, dass die acht Faktoren die Musik nicht vollständig beschreiben können, und dementsprechend große Störgrößen vorhanden sind.

Anpassung	**a** $R^2 = 0.13, R_a^2 = -0.09$		**b** $R^2 = 0.65, R_a^2 = 0.56$		**c** $R^2 = 0.61, R_a^2 = 0.51$	
Faktoren	Schätzer	p-Wert	Schätzer	p-Wert	Schätzer	p-Wert
(Intercept)	0.8448	**<2e-16**	0.6815	**<2e-16**	0.6945	**<2e-16**
Intervall	-0.0015	0.90	-0.0041	0.76	0.0308	0.17
Begl.-Einsätze	-0.0021	0.87	-0.0636	**4e-05**	-0.0448	**0.05**
Dynamik	-0.0146	0.25	-0.0186	0.17	-0.0019	0.93
Begl.-Instrument	0.0177	0.16	-0.0109	0.41	-0.1313	**2e-06**
Tonhöhe	0.0029	0.81	0.0510	**6e-04**	0.0198	0.37
Tondauer	-0.0087	0.49	-0.0348	**0.01**	-0.0026	0.91
Begl.-Tonhöhe	0.0051	0.68	-0.0224	**0.10**	-0.0213	0.34

Tabelle 7.2: Auswertung über alle Instrumente und alle Plackett-Burman-Designs für das beste Verfahren mit Ohrmodell (Aggregation der Einkanalschätzungen). Die Zielvariable ist der mittlere F-Wert – **a**: monophone Einsatzzeiterkennung, **b**: predominante Einsatzzeiterkennung und **c**: polyphone Einsatzzeiterkennung (**dick** = signifikant zum 10%-Level).

werden, die einen signifikanten Einfluss auf die Leistung des Algorithmus haben. Wie zu erwarten ist in der predominanten Variante die Leistung besser, wenn die Anzahl an individuellen Einsätzen der Begleitung gering ist. Aber auch höhere Tonhöhen und kürzere Töne verbessern die predominante Einsatzzeiterkennung. Da beide Effekte weder bei der monophonen noch bei der predominanten Variante zu beobachten sind, scheint sich beides positiv auf die Trennbarkeit von Einsätzen der Melodie und der Begleitung auszuwirken. Allerdings werden für die Einsatzzeiterkennung auch systematisch kürzere Tonlängen etwas bevorteilt. Denn durch die Maximumsbedingung verhindern richtig erkannte Toneinsätze falsche Schätzungen in ihrer Umgebung, und der relative Anteil dieser Umgebungen ist bei Musikstücken mit kurzen Tönen größer. Somit wird die falsche Erkennung von Toneinsätzen der Begleitung bei kürzeren Tonlängen stärker unterdrückt. Diese Feststellung gilt natürlich nur so lange sich die richtigen Toneinsätze der Melodie nicht gegenseitig unterdrücken, was bei sehr kurzen Tönen geschehen kann. Bei einer Mindesttonlänge von 0.1 s, die hier durch das Versuchsdesign gegeben ist, kann dies jedoch ausgeschlossen werden.

Bei der polyphonen Einsatzzeiterkennung erkennt das Verfahren die Einsätze in Musikstücken mit Klavierbegleitung besser als in Stücken mit Streicherbegleitung. Dies ist konsistent zu Erkenntnissen in anderen Studien, denn auf Grund von Frequenzmodulationen innerhalb eines Tons, sind Einsätze von Streichinstrumenten deutlich schwieriger zu identifizieren (Böck und Widmer, 2013). Auch die schlechteren Ergebnisse für das Cello,

die in Tabelle 7.1 zu sehen sind, sind darauf zurückzuführen. Des Weiteren scheint auch in dieser Variante ein kleinerer Anteil von individuellen Einsatzzeiten der Begleitung hilfreich. Zum einen sind gemeinsame Einsätze von Melodie und Begleitung deutlicher zu erkennen und zum anderen sind die Einzeltöne der Begleitung im Mittel etwas leiser.

7.1.2 Tonhöhenerkennung

In Tabelle 7.3 sind die durchschnittlichen Fehlerraten der Tonhöhenschätzung – unter Berücksichtigung einer Toleranz von einem halben Halbton ($\approx 3\%$ der Frequenz) – für die Verfahren, die in Kapitel 4 beschrieben sind, sowohl für den polyphonen als auch für die monophonen Datensatz, aufgelistet. Für den polyphonen Fall sind die Ergebnisse zudem bezüglich der drei untersuchten Musikinstrumente aufgeschlüsselt. Die besten Ergebnisse erzielt das Klassifikationsverfahren mit den spektralen Merkmalen und der linearen SVM mit einer Fehlerrate von 7% für polyphone Musik. Allerdings schneidet das gleiche Verfahren mit dem *Random Forest* nur minimal schlechter ab. Alle anderen Verfahren liefern deutlich schlechtere Ergebnisse, auch der YIN-Algorithmus, das Standardverfahren ohne Ohrmodell.[2] Beim SACF-Verfahren schneidet die Schwellenwertmethode, die kleinere Perioden bevorzugt, deutlich besser ab als die Standardmethode, die das globale Maximum auswählt. Ähnliche Ergebnisse liefern die Klassifikationsverfahren mit den ACF-Merkmalen. Warum diese Merkmale so viel schlechter abschneiden als die DFT-Merkmale, bleibt noch zu klären.

	polyphon (predominant)				monophon
Verfahren	Cello	Klarinette	Trompete	**Mittelwert**	Mittelwert
SACF maximaler Peak	0.55	0.52	0.54	0.54	0.20
SACF Schwellenwert	0.24	0.12	0.17	0.18	0.05
DFT-Merkmale + RF	0.14	0.02	0.08	0.08	0.02
DFT-Merkmale + SVM	0.11	0.01	0.08	**0.07**	0.02
ACF-Merkmale + RF	0.24	0.08	0.30	0.20	0.05
ACF-Merkmale + SVM	0.21	0.05	0.24	0.17	0.04
YIN-Algorithmus	0.36	0.15	0.32	0.28	0.05

Tabelle 7.3: Mittlere Fehlerraten der Verfahren zur Tonhöhenschätzung.

Für alle Varianten sind die Fehlerraten für die Klarinette mit Abstand die niedrigsten, wohingegen Cellotöne am schwierigsten zu erkennen sind. Bei Klarinettentönen haben die

[2] An dieser Stelle soll jedoch noch einmal angemerkt werden, dass der YIN-Algorithmus im Gegensatz zu den anderen Verfahren mit Standardeinstellungen läuft und nicht auf die Struktur der Daten angepasst ist.

ungeraden Obertöne nur schwache Intensitäten, so dass Oktavfehler unwahrscheinlicher werden. Hingegen wird bei Trompeten- oder Cellotönen oft die Frequenz des ersten Obertons (doppelte Grundfrequenz) fälschlicherweise als Grundfrequenz geschätzt. Die Erkennung der Cellotöne wird zudem durch die Ähnlichkeit zur Streicherbegleitung erschwert. Wie erwartet sind in der monophonen Variante die Fehlerraten aller Verfahren deutlich kleiner. Auch hier liefert die Klassifikation mit spektralen Merkmalen die besten Ergebnisse, wobei aber diesmal der *Random Forest* genauso gut wie die SVM abschneidet.

In der polyphonen Variante wird für die beste Methode – Klassifikationsansatz mit spektralen Merkmalen, sowohl mit SVM als auch mit *Random Forest* – eine gruppenbasierte Merkmalsselektion, wie in Kapitel 2.4 beschrieben, durchgeführt. Diese Ergebnisse sind in Tabelle 7.4 aufgelistet. Vor allem die Gruppierung auf Grund des Merkmalstyps ergibt eine starke Reduzierung der Merkmale bei weiterhin akzeptablen Fehlerraten. Für beide Klassifikationsmethoden endet die Vorwärtsvariante mit nur zwei Merkmalsgruppen mit nur leichten Einbußen im Vergleich zu den Ergebnissen mit 29 Merkmalsgruppen ohne Selektion. Interessanterweise werden für die zwei Klassifikationsverfahren jedoch unterschiedliche Merkmale ausgesucht: Für den *Random Forest* sind dies die Merkmale $c[k]$ und $d_{right}[k]$, das heißt die Information über die Grundfrequenzkandidaten und deren Abstände zu dem jeweiligen DFT-Maximum der höheren Frequenzen ist ausreichend, um eine Fehlerrate von 9% zu erreichen. Das gleiche Ergebnis liefert die lineare SVM mit den Merkmalen $p_{mean}[k]$ und $P_1^{mean}[k]$, das heißt mit der Information über die mittleren Feuerraten der Kanäle und über die mittleren Amplituden der Kandidaten über alle Kanäle. Auf Grund der identischen Fehlerraten, die zudem relativ nah an die Fehlerraten des vollständigen Modells heranreichen, lässt sich schließen, dass ein großer Teil der relevanten Information in beiden Merkmalspaaren redundant vorhanden ist, obwohl sie mathematisch vollkommen verschiedene Merkmale beschreiben.

Methode	ohne Sel.	Kanal-Gruppierung		Merkmalstyp-Gruppierung	
		vorwärts	rückwärts	vorwärts	rückwärts
RF	0.08	0.10	0.07	0.09	0.08
	$41 \cdot 29 = 1189$	$4 \cdot 29 = 116$	$35 \cdot 29 = 1015$	$41 \cdot 2 = 82$	$41 \cdot 28 = 1148$
SVM	0.07	0.10	0.07	0.09	0.07
	$41 \cdot 29 = 1189$	$5 \cdot 29 = 145$	$23 \cdot 29 = 667$	$41 \cdot 2 = 82$	$41 \cdot 9 = 369$

Tabelle 7.4: Merkmalsselektion für die Tonhöhenklassifikation mit DFT-Merkmalen der Ohrmodellausgabe: Fehlerraten und Anzahl der selektierten Merkmale.

In der Rückwärtsvariante benötigt die SVM nur neun Merkmalsgruppen um die gleichen Ergebnisse wie mit allen Merkmalen zu erreichen. Die ausgewählten Merkmale sind: $c[k]$, $p_{mean}[k]$, $p_{\max}[k]$, $b[k]$, $d_{left}[k]$, $d_{right}[k]$, $P_4^{mean}[k]$, $P_8^{mean}[k]$ und $P_9^{mean}[k]$. Bemerkenswert ist noch, dass die $P_{\text{pl}}^{\max}[k]$-Merkmale – die maximale Amplitude der Partialtöne über alle Kanäle – in keiner Variante ausgewählt werden.

Auch in der kanalbasierten Gruppierungsvariante können einige Kanäle unberücksichtigt bleiben. Für die SVM sind 23 Kanäle ausreichend, um gleich gute Ergebnisse (7% Fehlerrate) wie mit allen 41 Kanälen zu erzielen. Dabei sind die unberücksichtigten Kanäle in allen Bereichen verstreut, es kann also keine Präferenz zu höheren oder niedrigeren Kanälen erkannt werden. Ein Grund dafür ist die überlappende Struktur der Kanäle, wodurch ein Großteil der relevanten Information redundant in benachbarten Kanälen vorhanden ist.

Tabelle 7.5 zeigt die Auswertung des Versuchsplans für das beste Verfahren der Tonhöhenschätzung (Klassifikation mit DFT-Merkmalen und SVM). Die Anpassungsgüte des Modells ist mit $R_a^2 = 0.12$ eher niedrig, aber einige schwache Einflussfaktoren können identifiziert werden. Eine größere Distanz der Tonhöhen zwischen Melodie- und Begleitstimme scheint vorteilhaft zu sein. Dieser Effekt ist nicht unerwartet, da durch eine größere Tondifferenz, die Überschneidung der Töne im spektralen Bereich geringer wird. Des Weiteren hat die Art der Begleitung einen schwach signifikanten Effekt. Streicherbegleitung wirkt demnach stärker störend als Klavierbegleitung. Zumindest für die Melodietöne des Cellos ist dies auf Grund der Klangähnlichkeit zur Streicherbegleitung nachvollziehbar.

Anpassung	$R^2 = 0.30, R_a^2 = 0.12$	
Faktoren	Schätzer	p-Wert
(Intercept)	0.0660	**2e-08**
Intervall	-0.0074	0.40
Begl.-Einsätze	0.0056	0.53
Dynamik	-0.0142	0.11
Begl.-Instrument	0.0148	**0.10**
Tonhöhe	-0.0068	0.44
Tondauer	-0.0025	0.78
Begl.-Tonhöhe	-0.0185	**0.04**

Tabelle 7.5: Auswertung über alle Instrumente und alle Plackett-Burman-Designs für das Klassifikationsverfahren zur Tonhöhenschätzung mit DFT Merkmalen und linearer SVM. Die Fehlerrate ist die Zielvariable (**dick** = signifikant zum 10%-Level).

7.1.3 Instrumentenerkennung

Die Fehlerraten für die Instrumentenklassifikation sind in Tabelle 7.6 gelistet. Die Merkmale basierend auf dem Ohrmodell schneiden für die polyphonen Daten deutlich besser ab als die Standardmerkmale. In beiden Fällen liefert die lineare SVM etwas niedrigere Fehlklassifikationsraten ab als der *Random Forest*. Die beste Methode ist demnach das Klassifikationsverfahren, das die Ohrmodellmerkmale und die lineare SVM verwendet, mit einer Fehlklassifikationsrate von 1.1%. Trompete vom Rest zu trennen ist etwas schwieriger als Cello oder Klarinette zu identifizieren. In der monophonen Variante sind die Ergebnisse mit allen Methoden nahezu perfekt.

	polyphon (predominant)				monophon
Methode	Ce. vs. all	Kl. vs. all	Tr. vs. all	**Gesamt**	Gesamt
Ohrmodell, RF	0.012	0.017	0.029	0.019	0.002
Ohrmodell, SVM	0.007	0.007	0.014	**0.011**	0.001
Standard, RF	0.044	0.034	0.052	0.063	0.000
Standard, SVM	0.025	0.019	0.054	0.035	0.002

Tabelle 7.6: Mittlere Fehlklassifikationsraten der Instrumentenerkennung mit Ohrmodellmerkmalen oder Standardmerkmalen.

Abbildung 7.2 zeigt die Ergebnisse für die lineare SVM, wenn nur die Ohrmodellmerkmale eines Kanals verwendet werden. Je höher die Kanalnummer ist, desto niedriger ist die Fehlklassifikationsrate. Mit den Merkmalen des ersten Kanals wird eine Fehlerrate von fast 40% erreicht, wohingegen die Merkmale des 41. Kanals ein Modell mit einer Fehlerrate von unter 5% erzeugen. Dieses Ergebnis ist auch in Hinsicht auf Hörschädigungen interessant, bei denen meistens die höchsten Frequenzen (Kanäle) am stärksten beeinträchtigt sind. Auch in der Rückwärtsvariante der Kanal-basierten Gruppierung bleiben die niedrigsten Kanäle unberücksichtigt.

Tabelle 7.7 zeigt die Ergebnisse der Merkmalsselektion für die Instrumentenerkennung. Für den *Random Forest* werden die Ergebnisse durch beide gruppenbasierten Rückwärtsselektionsvarianten leicht verbessert. In der Kanal-basierten Variante reichen 12 Kanäle aus, um die gleiche niedrige Fehlerrate wie mit der linearen SVM mit allen Merkmalen zu erreichen. Die gewählten Kanäle sind 8, 12, 19, 21, 22, 24, 26, 27, 28, 32, 33 und 41, es werden also vor allem Kanäle mit mittleren Mittenfrequenzen ausgewählt. Die Merkmale der hohen Kanäle scheinen daher stark redundant zu sein. Die Ergebnisse der Vorwärtsselektion zeigen, dass zwei Kanäle ausreichend sind Fehlerraten von etwa 3% zu erreichen, womit eine akzeptable Variante mit niedriger Rechenzeit identifiziert wird. Für

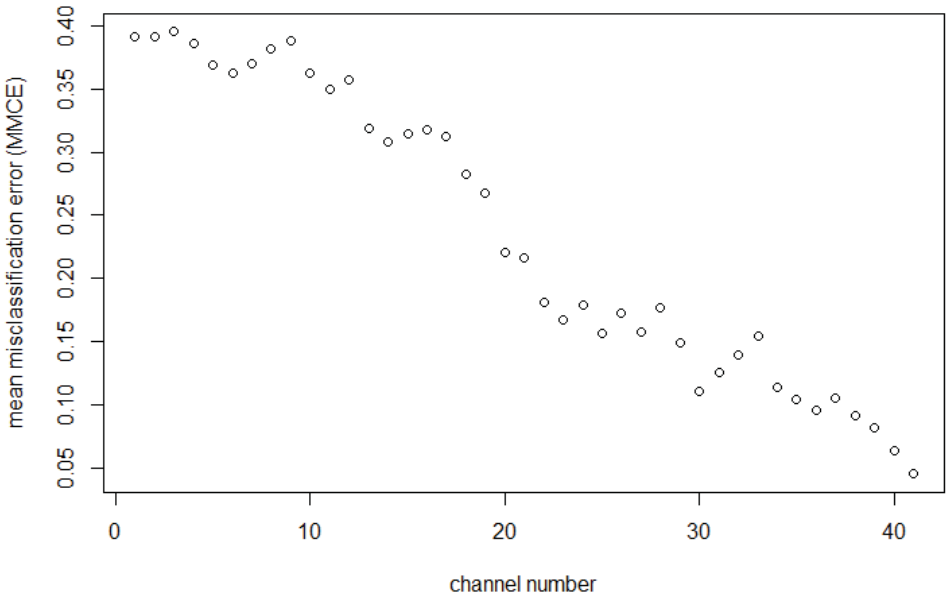

Abbildung 7.2: Fehlerraten für die predominante Instrumentenerkennung mit den Merkmalen nur eines Kanals und der linearen SVM.

den *Random Forest* werden die Kanäle 26 und 41 gewählt und für die SVM die Kanäle 29 und 41.

Methode	ohne Sel.	Kanal-Gruppierung		Merkmalstyp-Gruppierung	
		vorwärts	rückwärts	vorwärts	rückwärts
RF	0.019	0.034	0.011	0.058	0.016
	$41 \cdot 21 = 861$	$2 \cdot 21 = 42$	$12 \cdot 21 = 420$	$41 \cdot 3 = 123$	$41 \cdot 17 = 697$
SVM	0.011	0.030	0.017	0.045	0.015
	$41 \cdot 21 = 861$	$2 \cdot 21 = 42$	$12 \cdot 21 = 420$	$41 \cdot 3 = 123$	$41 \cdot 8 = 328$

Tabelle 7.7: Merkmalsselektion für die Instrumentenerkennung mit Ohrmodellmerkmalen: Fehlerraten und Anzahl der selektierten Merkmale.

Bei der Merkmalstyp-basierten Vorwärtsselektion werden die gleichen drei Merkmalstypen für die SVM und den *Random Forest* ausgewählt: *Mean Spectral Flux*, *RMS Energy* und *Spectral Rolloff 85*. In der Rückwärtsvariante werden für die SVM, neben diesen drei Merkmalen, zusätzlich noch die folgenden fünf Gruppen gewählt: *irregularity* und der erste, der dritte, der vierte und der siebte MFCC-Koeffizient.

Tabelle 7.8 zeigt die Auswertung des Versuchsplans für die predominante Instrumentenerkennung. Hier ist die Anpassungsgüte des Modells im Vergleich zu den anderen Erkennungsaufgaben moderat ($R_a^2 = 0.21$). Drei leicht signifikante Einflüsse werden identifiziert. Die höchste Signifikanz hat der Einfluss der Tonhöhe, tiefe Töne werden besser erkannt als hohe. Weiterhin hemmt Streichbegleitung die Erkennung mehr als Klavierbegleitung, wofür auch hier die Problematik der Klangähnlichkeit zwischen Cello und Streichbegleitung verantwortlich sein dürfte. Auch eine niedrige Distanz zwischen den Tonhöhen von Melodie- und Begleitstimme hemmt die Erkennung. In diesem Fall überschneiden sich mehr Partialtöne im Spektralbereich.

Anpassung	$R^2 = 0.37, R_a^2 = 0.21$	
Faktoren	Schätzer	p-Wert
(Intercept)	0.0111	**9e-04**
Intervall	0.0049	0.11
Begl.-Einsätze	0.0037	0.23
Dynamik	-0.0019	0.54
Begl.-Instrument	0.0056	**0.07**
Tonhöhe	0.0062	**0.05**
Tondauer	0.0025	0.42
Begl.-Tonhöhe	-0.0056	**0.07**

Tabelle 7.8: Auswertung über alle Instrumente und alle Plackett-Burman-Designs für die Instrumentenerkennung mit Ohrmodellmerkmalen und linearer SVM. Die Fehlerrate ist die Zielvariable (**dick** = signifikant zum 10%-Level).

7.2 Evaluierung der *Hearing Dummies*

In diesem Abschnitt werden die Ergebnisse der *Hearing Dummies* (HD), die in Kapitel 2.1.6 beschrieben sind, für die Einsatzzeiterkennung, die Tonhöhenschätzung und die Instrumentenklassifikation präsentiert. Im Vergleich zum Normalhörenden (NH) fallen bei all diesen *Hearing Dummies* Kanäle aus, wodurch Merkmale für die Klassifikationsverfahren fehlen. Bei *Hearing Dummy* 1 unterscheiden sich zudem auch die Ausgaben der verbliebenen Kanäle durch den Ausfall des linearen Pfades der DRNL-Filterbank.

	Monophone Einsatzzeiterkennung			
Hörschädigung	NH	HD1	HD2	HD3
Cello	0.79	0.67	0.74	0.78
Klarinette	0.76	0.75	0.77	0.72
Trompete	0.98	0.99	0.98	0.98
Mittelwert	0.84	0.80	0.83	0.83

	Predominante Einsatzzeiterkennung			
Hörschädigung	NH	HD1	HD2	HD3
Cello	0.46	0.37	0.44	0.45
Klarinette	0.72	0.58	0.70	0.69
Trompete	0.87	0.70	0.80	0.86
Mittelwert	0.68	0.55	0.65	0.67

	Polyphone Einsatzzeiterkennung			
Hörschädigung	NH	HD1	HD2	HD3
Cello	0.53	0.46	0.50	0.53
Klarinette	0.71	0.62	0.70	0.69
Trompete	0.85	0.74	0.81	0.83
Mittelwert	0.69	0.61	0.67	0.68

Tabelle 7.9: Ergebnisse (mittlere F-Werte) der Einsatzzeiterkennung für die *Hearing Dummies* (HD) im Vergleich zu dem Modell eines Normalhörenden (NH).

7.2.1 Einsatzzeiterkennung

Für die Einsatzzeiterkennung wird hier nur noch das beste Verfahren mit Ohrmodell betrachtet: die Aggregation der Einkanalschätzungen. Die Ergebnisse der Einsatzzeiterkennung für die drei *Hearing Dummies* sind in Tabelle 7.9 aufgelistet. In allen drei Varianten – monophon, predominant und polyphon – sind die Ergebnisse von HD2 und HD3 nur leicht schlechter als die Ergebnisse des normalen Ohrmodells ohne Hörschädigung. Dieses Ergebnis deutet darauf hin, dass die relativ moderaten Hörverluste keinen großen Einfluss auf die Wahrnehmung von Toneinsätzen haben, auch wenn die Ergebnisse der Einsatzzeiterkennung, auf Grund der insgesamt relativ hohen Fehlerraten, nur mit Vorsicht interpretiert werden sollten. Hier ist die Interpretation jedoch konsistent zu Erkenntnissen aus Hörversuchen, die die Rhythmuswahrnehmungen von Hörgeschädigten relativ positiv beurteilen (McDermott, 2004; Gfeller u. a., 2006; Looi u. a., 2008). Hingegen sind die Fehlerraten von HD1, der einen sehr starken Hörverlust nachbildet, deutlich höher, vor allem für die predominante Einsatzzeiterkennung.

Anpassung	**a** $R^2 = 0.50, R_a^2 = 0.38$		**b** $R^2 = 0.61, R_a^2 = 0.52$		**c** $R^2 = 0.81, R_a^2 = 0.76$	
Faktoren	Schätzer	p-Wert	Schätzer	p-Wert	Schätzer	p-Wert
(Intercept)	0.8011	**<2e-16**	0.5495	**<2e-16**	0.6072	**<2e-16**
Intervall	0.0020	0.86	0.0138	0.43	0.0429	**0.02**
Begl.-Einsätze	0.0054	0.65	-0.0438	**0.02**	-0.0169	0.32
Dynamik	-0.0197	0.11	0.0133	0.44	0.0082	0.63
Begl.-Instrument	0.0210	**0.09**	-0.0624	**8e-04**	-0.1705	**7e-11**
Tonhöhe	-0.0361	**5e-03**	0.0525	**5e-03**	0.0258	0.14
Tondauer	-0.0410	**2e-03**	-0.0424	**0.02**	-0.0205	0.23
Begl.-Tonhöhe	0.0064	0.59	-0.0453	**0.01**	-0.0353	**0.04**

Tabelle 7.10: Auswertung über alle Instrumente und alle Plackett-Burman-Designs für ***Hearing Dummy* 1**. Die Zielvariable ist der mittlere F-Wert – **a**: monophone Einsatzzeiterkennung, **b**: predominante Einsatzzeiterkennung und **c**: polyphone Einsatzzeiterkennung (**dick** = signifikant zum 10%-Level).

Die Auswertung des Versuchsplans für HD1, die in Tabelle 7.10 zu sehen ist, zeigt dann auch einige Unterschiede im Vergleich zum Normalhörenden auf (vergleiche Tabelle 7.2). Bei der monophonen Einsatzzeiterkennung werden die Einsätze tiefer Töne besser erkannt, ein Effekt der offensichtlich direkt der Hochfrequenzstörung zuzuordnen ist. Aber ein weiterer Einfluss, der nur für HD1, aber nicht für den Normalhörenden existiert, ist nicht so einfach zu erklären: Kürzere Töne werden besser erkannt als längere. Ein Grund dafür ist die schon angesprochene systematisch stärkere Unterdrückung von falschen Schätzungen bei Musikstücken mit kürzeren Tönen (vergleiche Kapitel 7.1.1). Die Verhinderung von falschen Schätzungen wirkt sich vor allem positiv in Situationen aus, in denen die Unsicherheit groß ist.

Bei der predominanten Einsatzzeiterkennung treten die meisten signifikanten Effekte auch für das Modell ohne Hörschädigung auf. Eine Ausnahme bildet jedoch das Begleitinstrument. Während für den Normalhörenden kein Einfluss zu erkennen ist, erzielt HD1 deutlich schlechtere $F-$ Werte bei Streicherbegleitung. In der predominanten Variante ist sowohl für HD1 als auch für das Modell ohne Hörschädigung die Art der Begleitung der dominierende Einflussfaktor. Beide Modelle haben deutlich mehr Probleme mit Streicherbegleitung als mit Klavierbegleitung. Daneben zeigt HD1 aber auch schwache signifikante Einflüsse für das Intervall (größere erleichtern die Identifizierung von Toneinsätzen) und für die mittlere Tonhöhe der Begleitung (ein kleinerer Tonhöhenabstand zur Begleitung ist vorteilhaft). Der letztgenannter Effekt ist überraschend, vor allem für die predominante Einsatzzeiterkennung, deren Ziel eine Trennung von Melodie und

Begleitung ist. Ein plausibler Grund dafür ist, dass bei einem höheren Tonhöhenabstand die Begleitung, bedingt durch den Versuchsaufbau, automatisch tiefer wird. Dadurch werden die Ausgaben der verbliebenen zehn Kanäle – alle mit tiefen *Best Frequenzen* – von der Begleitung dominiert, so dass die Toneinsätze der Melodie weniger hervorstechen.

Dagegen sind für HD2 und HD3 die Ergebnisse der Auswertung fast identisch zum Normalhörenden (siehe Tabellen A.5 und A.6 in Anhang A).

7.2.2 Tonhöhenerkennung

Für die Klassifikationsmethode zur Tonhöhenerkennung entsteht für HD1 und HD2 das Problem, dass durch den Ausfall von Kanälen nicht nur Merkmale wegfallen, sondern auch Klassenlabel verändert werden müssen. Beispielsweise bedeutet dies für HD1 – bei dem nur die Kanäle 1 bis 10 funktionieren –, dass Kanal 10 alle Tonhöhen identifizieren muss, für die sonst die höheren Kanäle zuständig sind. Dementsprechend werden für diese Töne die Klassenlabel auf „Kanal 10“ verändert.

In Tabelle 7.11 sind die Ergebnisse der predominanten Tonhöhenerkennung für die drei *Hearing Dummies* zusammengefasst. Für alle getesteten Methoden sind die Ergebnisse wie erwartet: Je größer der Hörverlust ist, desto höher sind auch die Fehlerraten. Auch für alle *Hearing Dummies* liefert die Klassifikationsmethode mit DFT-Merkmalen und linearer SVM die besten Ergebnisse. Selbst HD3 schneidet etwas schlechter als NH ab, obwohl dessen Hördefizit nicht die Grundfrequenzen der berücksichtigten Tonhöhen betrifft. Dieses Ergebnis ist jedoch konsistent zu Ergebnissen aus psychoakustischen Experimenten, die darauf hindeuten, dass auch die Obertöne einen wichtigen Einfluss auf die Tonhöhenerkennung haben (Oxenham, 2008).

Methode	NH	HD1	HD2	HD3
SACF maximaler Peak	0.54	0.67	0.60	0.56
SACF Schwellenwert	0.18	0.44	0.34	0.22
DFT-Merkmale + RF	0.08	0.32	0.29	0.10
DFT-Merkmale + SVM	0.07	0.32	0.24	0.09
ACF-Merkmale + RF	0.20	0.91	0.42	0.21
ACF-Merkmale + SVM	0.17	0.90	0.40	0.18

Tabelle 7.11: Mittlere Fehlerraten der Tonhöhenschätzung für die *Hearing Dummies* (HD) im Vergleich zum Normalhörenden (NH).

Tabelle 7.12 listet die Ergebnisse der Merkmalsselektion auf. Für alle *Hearing Dummies* fällt auf, dass, trotz einer erhebliche Reduzierung der Kanal- oder der Merkmalstyp-Gruppen, fast identische Ergebnisse wie mit dem vollständigen Modell erzielt werden. Demzufolge sind viele der Merkmale bedeutungslos oder redundant. Dies mag an der Struktur der Merkmale liegen, die im Vergleich zur Instrumentenerkennung nicht vollständig unabhängig aus den einzelnen Kanälen extrahiert werden. Die Merkmale $P_{\mathrm{pl}}^{\mathrm{mean}}[k]$ und $P_{\mathrm{pl}}^{\mathrm{max}}[k]$ sind nämlich über alle Kanäle definiert.

Methode	ohne Selektion	Kanal-Gruppierung		Merkmalstyp-Gruppierung	
		vorwärts	rückwärts	vorwärts	rückwärts
HD1	0.32	0.32	0.32	0.32	0.32
RF	$10 \cdot 29 = 290$	$2 \cdot 29 = 58$	$7 \cdot 29 = 203$	$10 \cdot 2 = 20$	$10 \cdot 15 = 150$
HD1	0.32	0.33	0.31	0.32	0.32
SVM	$10 \cdot 29 = 290$	$1 \cdot 29 = 29$	$8 \cdot 29 = 232$	$10 \cdot 1 = 10$	$10 \cdot 9 = 90$
HD2	0.29	0.33	0.29	0.28	0.29
RF	$25 \cdot 29 = 725$	$2 \cdot 29 = 58$	$24 \cdot 29 = 696$	$25 \cdot 3 = 75$	$25 \cdot 26 = 650$
HD2	0.24	0.26	0.25	0.27	0.24
SVM	$25 \cdot 29 = 725$	$7 \cdot 29 = 203$	$22 \cdot 29 = 638$	$25 \cdot 3 = 75$	$25 \cdot 21 = 525$
HD3	0.10	0.10	0.10	0.10	0.10
RF	$29 \cdot 29 = 841$	$4 \cdot 29 = 116$	$27 \cdot 29 = 783$	$29 \cdot 2 = 58$	$29 \cdot 24 = 696$
HD3	0.09	0.10	0.09	0.10	0.10
SVM	$29 \cdot 29 = 841$	$6 \cdot 29 = 174$	$23 \cdot 29 = 667$	$29 \cdot 3 = 87$	$29 \cdot 12 = 348$

Tabelle 7.12: Merkmalsselektion für die Tonhöhenklassifikation mit DFT-Merkmalen für die *Hearing Dummies*: Fehlerraten und Anzahl der selektierten Merkmale.

Die Auswertung des Versuchsplans für die *Hearing Dummies* ist in Tabelle 7.13 zu sehen. Hier wird nur die Klassifikationsmethode mit DFT-Merkmalen und linearer SVM betrachtet. Die Ergebnisse von HD3 sind relativ ähnlich zu den Ergebnissen des Normalhörenden (vergleiche Tabelle 7.5). Für diesen wird kein signifikanter Einfluss gefunden, das heißt die Tonhöhenerkennung ist für jeden Musiktyp gleich schwierig (sofern keine Wechselwirkungen vorliegen). Dagegen sind für HD1 und HD2 deutlich größere Unterschiede zu erkennen. Für HD2 sind hohe Töne problematisch, was konsistent zu der Struktur des Hördefizits ist. Dagegen hat HD1 überraschend Probleme bei der Tonhöhenidentifizierung von tiefen Tönen. Grund dafür ist ein Artefakt des Klassifikationsmodells bedingt durch die ungleiche Klassenverteilung – „Kanal 10“ ist mit Abstand die häufigste Klasse – und die schlechte Trennbarkeit der Daten auf Grund der hörschädigungsbedingten Merkmalsreduktion und -verfälschung. Das hat zur Folge, dass auch für tiefe Tonhöhen oft

fälschlicherweise der zu hohe Tonhöhenkandidat von Kanal 10 gewählt wird. Für HD1 hat zudem auch die Dynamik einen signifikanten Einfluss, wobei ein hoher Wert die Identifikation erleichtert. Ein schwach signifikanten Effekt zeigt sich noch für HD2 und die Tondauer, der darauf hindeutet, dass die Tonhöhe längerer Töne leichter zu schätzen ist. Dieser Effekt ist nicht ungewöhnlich, auch wenn hier die minimale Tondauer auf 0.1 s festgelegt ist (vergleiche Kapitel 4.2.1).

	HD1		**HD2**		**HD3**	
Anpassung	$R^2 = 0.58, R_a^2 = 0.47$		$R^2 = 0.37, R_a^2 = 0.21$		$R^2 = 0.20, R_a^2 = 0.00$	
Faktoren	Schätzer	p-Wert	Schätzer	p-Wert	Schätzer	p-Wert
(Intercept)	0.3154	**<2e-16**	0.2441	**<2e-16**	0.0923	**<2e-16**
Intervall	-0.0278	0.12	-0.0083	0.72	-0.0034	0.74
Begl.-Einsätze	-0.0327	**0.07**	0.0114	0.62	0.0102	0.32
Dynamik	-0.0630	**1e-03**	-0.0386	0.10	-0.0102	0.32
Begl.-Instrument	0.0154	0.38	0.0133	0.57	0.1120	0.24
Tonhöhe	-0.0679	**5e-04**	0.0645	**9e-03**	-0.0009	0.93
Tondauer	-0.0228	0.19	-0.0503	**0.04**	-0.0077	0.45
Begl.-Tonhöhe	0.0080	0.64	0.0040	0.86	-0.0170	0.10

Tabelle 7.13: Auswertung über alle Instrumente und alle Plackett-Burman-Designs für die Tonhöhenklassifikation der *Hearing Dummies* (HD) mit DFT-Merkmalen und SVM. Die Fehlerrate ist die Zielvariable – (**dick** = signifikant zum 10%-Level).

7.2.3 Instrumentenerkennung

Auch für die Instrumentenklassifikation sind die Ergebnisse konsistent zum Grad der Schwerhörigkeit, wie Tabelle 7.14 zeigt. Hier schneidet jedoch diesmal auf Grund der Wichtigkeit der hohen Kanäle (vergleiche Abbildung 7.2) HD2 besser ab als HD3.

Methode	NH	HD1	HD2	HD3
Ohrmodellmerkmale, RF	0.02	0.28	0.03	0.05
Ohrmodellmerkmale, SVM	0.01	0.26	0.02	0.04

Tabelle 7.14: Mittlere Fehlklassifikationsraten der Instrumentenerkennung für die *Hearing Dummies* (HD) im Vergleich zum Normalhörenden (NH).

Die Merkmalsselektion der Instrumentenerkennung ist insbesondere für eine Beschleunigung der Hörgeräteoptimierung wichtig (siehe Kapitel 8.2), da die Reduktion der Merkmale eine erhebliche Reduktion der Rechenzeit mit sich bringt. Die Ergebnisse für

die *Hearing Dummies* sind in Tabelle 7.15 dargestellt. Ähnlich wie bei der Tonhöhenerkennung, scheint es auch hier eine Vielzahl redundanter Merkmale zu geben. Möglich ist aber auch, dass die verwendeten Klassifikationsmodelle Probleme mit der hohen Anzahl von Merkmalen des vollständigen Modells haben. Dafür spricht, dass die lineare SVM für HD3 mit 14 Kanal-Gruppen besser abschneidet als mit allen 21 (3% Fehlerrate gegenüber 4%).

Methode	ohne Selektion	Kanal-Gruppierung		Merkmalstyp-Gruppierung	
		vorwärts	rückwärts	vorwärts	rückwärts
HD1	0.28	0.27	0.27	0.31	0.28
RF	$10 \cdot 21 = 210$	$2 \cdot 21 = 42$	$4 \cdot 21 = 84$	$10 \cdot 6 = 60$	$10 \cdot 17 = 170$
HD1	0.26	0.26	0.26	0.29	0.27
SVM	$10 \cdot 21 = 210$	$3 \cdot 21 = 63$	$6 \cdot 21 = 126$	$10 \cdot 7 = 70$	$10 \cdot 19 = 190$
HD2	0.03	0.04	0.03	0.05	0.02
RF	$25 \cdot 21 = 525$	$2 \cdot 21 = 42$	$22 \cdot 21 = 462$	$25 \cdot 4 = 100$	$25 \cdot 25 = 625$
HD2	0.02	0.04	0.02	0.06	0.02
SVM	$25 \cdot 21 = 525$	$2 \cdot 21 = 42$	$9 \cdot 21 = 189$	$25 \cdot 4 = 100$	$25 \cdot 12 = 300$
HD3	0.05	0.06	0.05	0.08	0.05
RF	$29 \cdot 21 = 609$	$2 \cdot 21 = 42$	$28 \cdot 21 = 588$	$29 \cdot 3 = 87$	$29 \cdot 14 = 406$
HD3	0.04	0.05	0.03	0.08	0.04
SVM	$29 \cdot 21 = 609$	$4 \cdot 21 = 84$	$14 \cdot 21 = 294$	$29 \cdot 4 = 116$	$29 \cdot 17 = 493$

Tabelle 7.15: Merkmalsselektion für die Instrumentenklassifikation mit Ohrmodellmerkmalen für die *Hearing Dummies*: Fehlerraten und Anzahl der selektierten Merkmale.

Die Ergebnisse der Versuchsplanauswertung für die Instrumentenerkennung, die in Tabelle 7.16 aufgelistet sind, sind sehr ähnlich zu den Ergebnissen des Normalhörenden. Für alle drei *Hearing Dummies* sind Musikinstrumente bei tiefen Tonhöhen und Klavierbegleitung einfacher zu identifizieren. Des Weiteren zeigt sich für HD2 und HD3 noch ein signifikanter Effekt bei der mittleren Tonhöhe der Begleitung, wonach ein größerer Abstand zur Melodie die Instrumentenerkennung erleichtert. Dagegen ist dieser Effekt bei HD1 nicht zu erkennen.

Anpassung	**HD1** $R^2 = 0.39, R_a^2 = 0.24$		**HD2** $R^2 = 0.44, R_a^2 = 0.29$		**HD3** $R^2 = 0.61, R_a^2 = 0.52$	
Faktoren	Schätzer	p-Wert	Schätzer	p-Wert	Schätzer	p-Wert
(Intercept)	0.2559	**<2e-16**	0.0210	**1e-05**	0.0373	**4e-09**
Intervall	0.0096	0.43	0.0080	**0.05**	0.0090	0.05
Begl.-Einsätze	0.0176	0.15	0.0056	0.17	0.0090	0.05
Dynamik	-0.0028	0.82	0.0012	0.76	-0.0046	0.31
Begl.-Instrument	0.0380	**3e-03**	0.0080	**0.05**	0.0145	**3e-03**
Tonhöhe	0.0213	**0.08**	0.0099	**0.02**	0.0182	**3e-04**
Tondauer	0.0139	0.25	0.0025	0.54	0.0040	0.38
Begl.-Tonhöhe	-0.0022	0.86	-0.0086	**0.04**	-0.0120	**0.01**

Tabelle 7.16: Auswertung über alle Instrumente und alle Plackett-Burman-Designs für die Instrumentenerkennung der *Hearing Dummies* (HD) mit SVM. Die Fehlerrate ist die Zielvariable – (**dick** = signifikant zum 10%-Level).

8 Hörgeräteoptimierung

In diesem Kapitel wird ein Hörgerätealgorithmus beispielhaft für eine starke Hörschädigung (*Hearing Dummy* 1, siehe Kapitel 2.1.6)) durchgeführt. Für dieses Beispiel wird jedoch für die Zielfunktion lediglich die Fehlklassifikationsrate der Instrumentenerkennung berücksichtigt. Auf die Einsatzzeiterkennung und die Tonhöhenschätzung wird auf Grund des hohen Rechenzeitbedarfs verzichtet. Zunächst wird in Abschnitt 8.1 der zu optimierende Hörgerätealgorithmus vorgestellt. Der Aufbau der Optimierung wird anschließend in Abschnitt 8.2 beschrieben. Dort werden auch die Ergebnisse präsentiert. Am Ende dieses Kapitels wird in Abschnitt 8.3 diskutiert, wie eine vollständige Optimierung, die auch die Einsatzzeiterkennung und die Tonhöhenschätzung berücksichtigt, ablaufen kann. Dafür ist eine Aggregation der individuellen Fehlermaße notwendig. Zudem werden Möglichkeiten zur Rechenzeitreduzierung der Optimierung aufgezeigt.

8.1 Hörgerätealgorithmus

Der zu optimierende Algorithmus ist in Kooperation mit einem Projektpartner im Rahmen eines gemeinsamen Forschungsprojekts[1] entwickelt worden. Die grundlegende Idee des Verfahrens ist es, die spektrale Komplexität eines Musiksignals zu reduzieren, um somit die Hörwahrnehmung von Musik für hörgeschädigte Personen zu vereinfachen (Nagathil, Weihs und Martin, 2016). Der Aufbau des Verfahrens ist in Abbildung 8.1 skizziert.

Zunächst werden die Toneinsatzzeiten der Melodiestimme mit Hilfe des in Kapitel 3.1 beschriebenen Verfahrens geschätzt. Da an dieser Stelle noch nicht die Verwendung eines Ohrmodells nötig ist, wird das Standardverfahren verwendet, das die Schätzung direkt auf der akustischen Wellenform durchführt. Die gefundenen Einsatzzeitpunkte definieren für die spätere Komplexitätsreduktion die zeitlichen Blöcke, die gemeinsam betrachtet werden. Anschließend wird die Constant-Q-Transformation (CQT) durchgeführt, die

[1] Teilprojekt B3 des SFB 823: „Statistische Modellierung zeitlich und spektral hoch aufgelöster Audiodaten in Hörgeräten“, Projektpartner: Institut für Kommunikationsakustik der Ruhr-Universität Bochum.

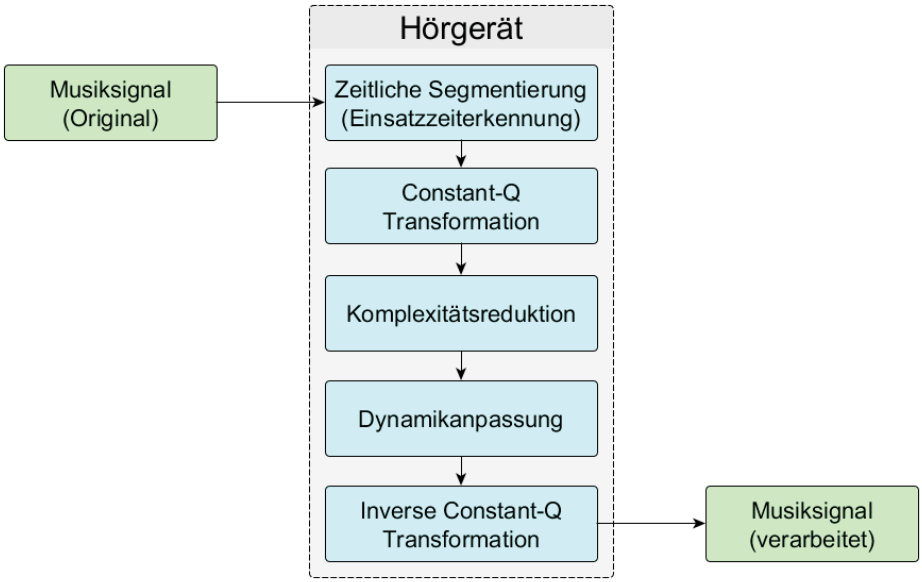

Abbildung 8.1: Hörgeräteaufbau.

ähnlich wie die Kurzzeit-Fourier-Transformation (STFT, siehe Gleichung 3.2) das Signal in den Frequenzraum transformiert, wobei jedoch die Frequenzlinien an die Notenskala der westlichen Musik angepasst sind. Dies führt zu höheren Bandweiten für die höheren Frequenzen, während bei der STFT alle Bandweiten gleich groß sind. Entsprechend werden, statt konstanter Segmentlängen wie bei der STFT, variable Segmentlängen verwendet. Es besteht ein antiproportionaler Zusammenhang zwischen den frequenzabhängigen Fensterlängen und den Bandbreiten. Je länger das Fenster (tiefe Frequenzen), desto schmaler werden die Abstände der Frequenzlinien. Die CQT ist durch

$$X_{\mathrm{cqt}}[n,\mu] = \frac{1}{M_\mu} \sum_{k=0}^{M_\mu - 1} x[h(n)+k] w_\mu[k] \exp\left(-i \frac{2\pi Q k}{M_\mu}\right) \tag{8.1}$$

definiert. Dabei bezeichnet M_μ die Anzahl der CQT-Koeffizienten, $h(n)$ den ersten Abtastzeitpunkt von Segment n und $w_\mu[k]$ eine Fensterfunktion. $Q = 1/(2^{\frac{1}{12b}} - 1)$ ist ein konstanter Gütefaktor, der das Verhältnis zwischen den Frequenzhöhen und den Abständen der Frequenzlinien beschreibt. In der Standardeinstellung mit $b = 1$ wird jede Oktave entsprechend der Notenskala in 12 CQT-Frequenzlinien eingeteilt.

Der nächste Schritt ist die Komplexitätsreduktion, die unabhängig für jeden durch die Einsatzzeiterkennung definierten zeitlichen Block durchgeführt wird. Dafür wird zunächst

auf den CQT-transformierten Segmenten eine Hauptkomponentenanalyse durchgeführt. Anschließend wird eine Niedrig-Rang-Approximation des Blocks erzielt, indem nur eine vorher festgelegte Anzahl von Komponenten beibehalten wird, die entsprechend ihrer Eigenwerte ausgewählt werden. Der folgende Schritt ist eine Dynamikkompression, die ein Musiksignal an den individuellen wahrnehmbaren Dynamikbereich einer hörgeschädigten Person anpasst. Dabei wird jede Frequenz unabhängig behandelt.

Die Dynamikkompression ist abhängig von der Ansprechzeit at, der Rücklaufzeit rt und einem frequenzabhängigen Verstärkungsfaktor. Die Ansprechzeit ist die Verzögerungszeit, bis die Dynamikkompression auf Pegelanstiege des Signals reagiert. Die Rücklaufzeit definiert die Zeit, wie lange höhere Pegel der Vergangenheit im Gedächtnis bleiben. Beide Zeiten sind Parameter, die optimiert werden müssen. Für die Berechnung der frequenzabhängigen Verstärkung wird zunächst der Schalldruckpegel jeder Frequenzlinie μ zum Zeitpunkt (Segmentnummer) n durch

$$X_{\text{spl}}[n,\mu] = 20 \cdot \log_{10}(|X_{\text{cqt}}[n,\mu]|) \tag{8.2}$$

ermittelt. Anschließend wird der digitale Ist-Pegel $P_{\text{ist}}[n,\mu]$ unter Berücksichtigung der Schätzung des vorherigen Zeitpunkts geschätzt:

$$P_{ist}[n,\mu] = \begin{cases} AF \cdot X_{\text{spl}}[n,\mu] + (1-AF) \cdot P_{\text{ist}}[n-1,\mu], & \text{falls } X_{\text{spl}}[n,\mu] > P_{\text{ist}}[n-1,\mu] \\ RF \cdot X_{\text{spl}}[n,\mu] + (1-RF) \cdot P_{\text{ist}}[n-1,\mu], & sonst, \end{cases} \tag{8.3}$$

mit dem Ansprechfaktor AF und dem Rücklauffaktor RF, die durch

$$\begin{aligned} AF &= 1 - \exp\left(\frac{-80 * at}{sF}\right) \text{ und} \\ RF &= 1 - \exp\left(\frac{-80 * rt}{sF}\right) \end{aligned} \tag{8.4}$$

definiert sind, wobei $sF = 44100$ Hz die Abtastrate bezeichnet. Der Verstärkungsfaktor ist dann abhängig von der Differenz des Ist-Pegels zu dem gewünschten Soll-Pegel $P_{soll}[n,\mu]$, der durch

$$P_{\text{soll}}[n,\mu] = \begin{cases} P_{\text{ist}}[n,\mu], & \text{falls } P_{\text{ist}}[\mu] < -90 \\ P_{\text{ist}}[n,\mu] + T[\mu] + \min\left(0, \frac{T[\mu]\cdot(P_{ist}[\mu]-KP_T)}{5+KP_T}\right), & sonst \end{cases} \tag{8.5}$$

definiert ist. Dabei bezeichnet $T[\mu]$ eine frequenzabhängige maximalen Verstärkung,

und $KP_T = min(KP, (-5 - T_\mu))$ ist der Kniepunkt, ab dem der Grad der Verstärkung abgeschwächt wird. KP ist ein zu optimierender Parameter. Ein höherer Wert von KP bewirkt, dass auch höhere Pegel maximal verstärkt werden. Allerdings ist durch die Gleichung auch der maximal zu erreichende Soll-Pegel auf -5dB beschränkt (siehe auch Abbildung 8.2), der bei einem höheren Wert von KP auch für niedrigere Pegel erreicht wird. Die Deckelung des Soll-Pegels verhindert, dass die sogenannte Unbehaglichkeitsschwelle überschritten wird. Für den Hörer sind jedoch alle Pegel, die auf diesen Wert angehoben werden, nicht mehr unterscheidbar.

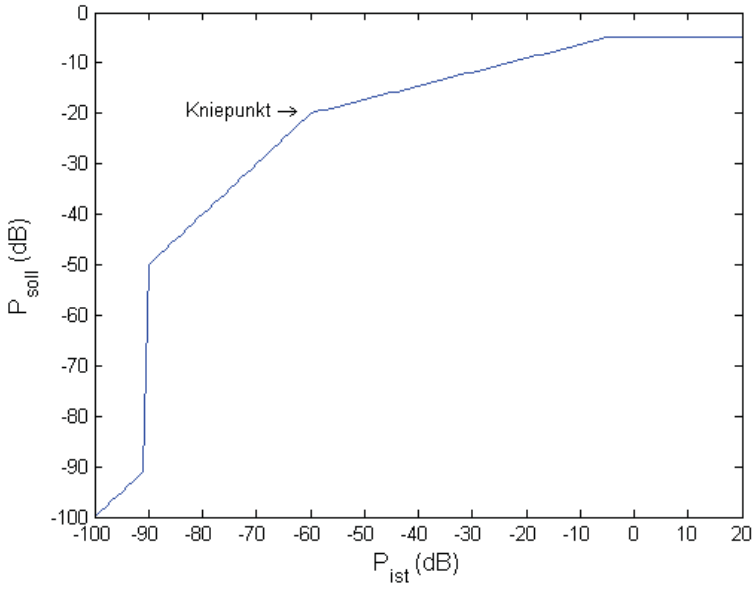

Abbildung 8.2: Zusammenhang zwischen vorliegenden Pegel P_{ist} und gewünschtem Pegel P_{soll}, beispielhaft für $T[\mu] = 40$ und KP = -60.

$T[\mu]$ ist ein Parameter, der für jeden Hörgeschädigten (und jedes $[\mu]$) individuell festgelegt werden muss. Für *Hearing Dummy* 1 wurde hierfür eine Experteneinstellung[2], die in Abbildung 8.3 visualisiert ist, verwendet. Prinzipiell könnte man $T[\mu]$ auch optimieren, dies würde aber die Dimension des Optimierungsproblems stark erhöhen. Der Zusammenhang zwischen P_{ist} und P_{soll} ist beispielhaft für $T[\mu] = 40$ und KP = -60 in Abbildung 8.2 dargestellt.

Im letzten Schritt werden die spektral reduzierten Segmente mittels eines approximativen Verfahrens – es existiert keine Inverse der CQT – wieder in den Zeitbereich zurück-

[2]Institut für Kommunikationsakustik der Ruhr-Universität Bochum

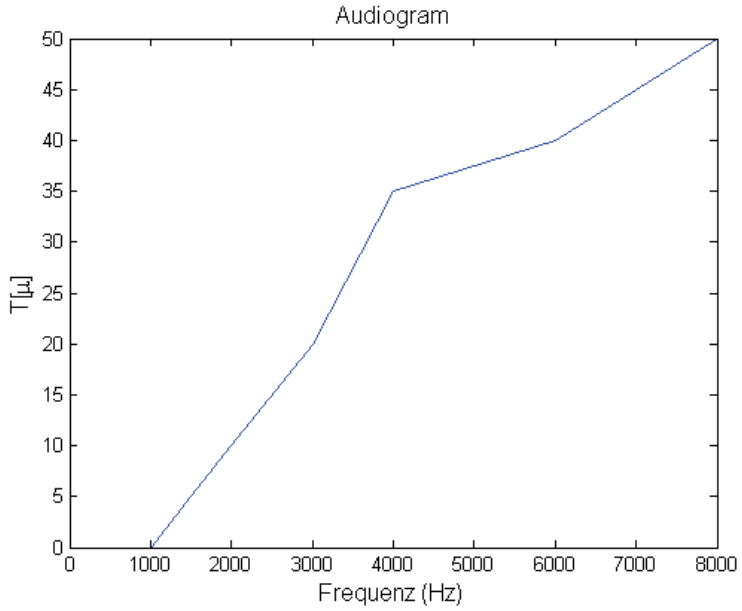

Abbildung 8.3: Maximale Verstärkung $T[\mu]$ für *Hearing Dummy* 1 (frequenzabhängig).

transformiert und durch ein Syntheseverfahren zusammengefügt (Nagathil und Martin, 2012).

8.2 Optimierungsexperiment

Die Optimierung soll beispielhaft für *Hearing Dummy* 1 (HD1) mit der Fehlklassifikationsrate der Instrumentenerkennung als Zielvariable durchgeführt werden. Der Aufbau der Optimierung ist in Abbildung 8.4 schematisch dargestellt. Für die Bewertung wird die Instrumentenerkennung gewählt, da diese kostengünstig, mit nur zwei Kanälen (Kanäle 5 und 10) des Ohrmodells, approximiert werden kann (Fehlklassifikation von 27% gegenüber 26% bei Verwendung aller zehn Kanäle von HD1, siehe Tabelle 7.15). Zudem existiert für die Instrumentenerkennung ein großes Verbesserungspotential, denn für das Ohrmodell ohne Hörschädigung beträgt die Fehlklassifikationsrate bei einer Verwendung von 2 Kanälen nur 3% (siehe Tabelle 7.7).

Insgesamt werden elf Parameter des Hörgerätealgorithmus optimiert, von denen sich sieben auf die Einsatzzeiterkennung beziehen, einer auf die Komplexitätsreduktion (Anzahl der PCA-Komponenten) und drei auf die Dynamikkompression (Ansprechzeit, Rücklaufzeit

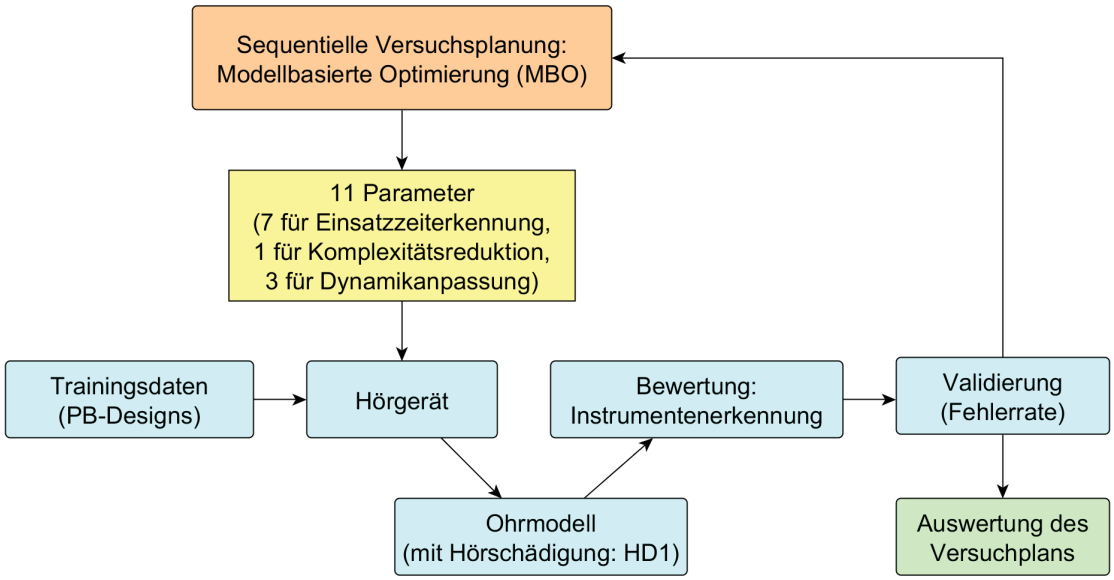

Abbildung 8.4: Optimierungsaufbau.

und der Kniepunkt KP). Im Gegensatz zu der Optimierung der Einsatzzeiterkennung in Kapitel 3.2, wird zur Vereinfachung der Optimierung auf vier weniger wichtige Parameter verzichtet, die statt dessen auf folgende Werte festgesetzt werden: $\gamma = 5$, $\alpha = 0.9$, $l_T = 0.3$ s und $r_T = 0.3$ s. Die zu optimierenden Variablen sind mit ihren zugehörigen Definitionsbereichen in Tabelle 8.1 zusammengefasst.

Es werden die gleichen Daten verwendet, die durch das Versuchsdesign in Kapitel 6 definiert sind. Somit kann hinterher auch wieder der Versuchsplan ausgewertet werden. Das heißt es gibt drei Mengen mit je 36 Musikstücken (Instanzen), die mit Hilfe von PB-Designs ausgewählt werden. Für die Bewertung einer Parameterkonfiguration wird auch hier eine dreifache Kreuzvalidierung angewendet, indem jeweils zwei Designs die Trainings- und das dritte Design die Testmenge bilden. Zunächst werden alle Musikdaten durch das entsprechend konfigurierte Hörgerät transformiert (vergleiche Abbildung 8.1). Anschließend werden die modifizierten Musiksignale durch das Ohrmodell mit Hörschädigung in simulierte neuronale Aktivitätswahrscheinlichkeiten umgewandelt, aus denen dann die Merkmale für die Instrumentenerkennung extrahiert werden. Mit den Beobachtungen der Trainingsmenge und diesen Merkmalen wird ein Klassifikationsmodell (*Random Forest*) angepasst, dessen Güte schließlich mit Hilfe der Testmenge evaluiert wird. Die mittleren Fehlerraten der drei Iterationen der Kreuzvalidierung ergeben die Kosten der Parameterkonfiguration.

Komponente des Hörgeräts	Parameter	Minimaler Wert	Maximaler Wert
Einsatzzeiterkennung	Fenstergröße M	1024	4096
	Sprungweite h	400	1600
	λ	0	1
	δ	0	10
	l_O	0 s	0.25 s
	r_O	0 s	0.25 s
	Verschiebung τ	-0.025 s	0.025 s
Komplexitätsreduktion	$\#Komponenten$	1	30
Dynamikkompression	Ansprechzeit at	0 s	3 s
	Rücklaufzeit rt	0 s	3 s
	Kniepunkt KP	-100	-10

Tabelle 8.1: Parameter und ihre Definitionsbereiche für die Hörgeräteoptimierung.

Trotz aller Vereinfachungen ist die Optimierung immer noch sehr zeitaufwändig. Wenn eine neue Parameterkonfiguration des Hörgerätealgorithmus getestet werden soll, muss für jedes betrachtete Musikstück das Hörgerät, das Ohrmodell und die Merkmalsextraktion erneut angewendet werden. Zum Schluss müssen noch die Klassifikationsmodelle erstellt und evaluiert werden. Vor allem die Ohrmodelltransformation und die Merkmalsextraktion benötigen eine sehr hohe Rechenzeit. Auf dem verwendeten Computercluster[3] werden beispielsweise für die 108 betrachteten Musikstücke ca. drei Stunden für die Ohrmodelltransformation eines Kanals und nochmal die gleiche Zeit für die Extraktion der Merkmale (auch nur für einen Kanal) benötigt. Wenn man diese Zahlen auf das vollständige Ohrmodell mit 41 Kanälen hochrechnet, bedeutet dies eine Rechenzeit von zehn Tagen für die Auswertung einer einzigen Parameterkonfiguration. Neben der Kanalreduzierung, wird dieser Zeitbedarf auch durch eine interne Parallelisierung vermindert, durch die bis zu zwölf Musikstücke gleichzeitig verarbeitet werden. Der Zeitbedarf für die Auswertung einer Parameterkonfiguration kann durch diese beiden Maßnahmen auf etwa eine Stunde reduziert werden.

Die Parameter werden mit MBO mit 275 Iterationen optimiert, wobei 55 für das LHS-Startdesign und 220 für die Iterationsschritte verwendet werden. Das optimierte Ergebnis wird zum einen mit der Fehlklassifikationsrate von *Hearing Dummy* 1 ohne Hörgerät verglichen (27%). Des Weiteren existiert noch eine Experteneinstellung[4], bei der statt der Einsatzzeiterkennung konstante Segmentlängen (0.1 s.) verwendet werden und die übrigen Parameter wie folgt gesetzt wurden: $\#Komponenten = 10$, $at = 2s$, $rt = 2s$

[3] *High Performance Computer-Cluster LiDOng.*
[4] Institut für Kommunikationsakustik der Ruhr-Universität Bochum

und $KP = -50$. Mit diesen Einstellungen ergibt sich eine Fehlklassifikationsrate von 19%. Durch die Optimierung kann die Fehlerrate jedoch noch weiter auf 14% gesenkt werden, womit diese, im Vergleich zur Variante ohne Hörgerät, fast halbiert ist. Die dazugehörige Parameterkonfiguration ist $M = 2551$, $h = 477$, $\lambda = 0.35$, $\delta = 2.73$, $l_O = 0.16$ s, $r_O = 0.11$ s, $\tau = -0.002$ s, $\#Komponenten = 5$, $at = 1.47$ s, $rt = 0.31$ s und $KP = -22.59$. Im Vergleich zu der Experteneinstellung ist die Anzahl der verwendeten Komponenten deutlich niedriger (5 statt 10). Außerdem werden von der Optimierung kleinere Werte für die Ansprech- und die Rücklaufzeit gewählt. Der Kniepunkt wird relativ hoch gewählt (auch moderate Pegel werden somit stark verstärkt), was für die starke Hörschädigung jedoch auch plausibel erscheint. Die aus dem optimalen Kniepunkt $KP = -22.59$ abgeleiteten Kennlinien sind in Abbildung B.2 im Anhang B visualisiert.

Die Auswertung des Versuchsplans für das optimierte Hörgerät ist in Tabelle 8.2 zu sehen. Kein Faktor hat einen signifikanten Einfluss. Aus dem Vergleich zu den Ergebnissen ohne Hörgerät, bei denen das Begleitinstrument einen signifikanten Einfluss hat (siehe Tabelle 7.16), lässt sich schließen, dass das optimierte Hörgerät vor allem die Klassifikationsergebnisse für die Musikstücke mit Streicherbegleitung verbessert.

Anpassung	$R^2 = 0.12, R_a^2 = -0.10$	
Faktoren	Schätzer	p-Wert
(Intercept)	0.1407	**1.37e-13**
Intervall	0.0130	0.23
Begl.-Einsätze	0.0124	0.26
Dynamik	6e-18	1.00
Begl.-Instrument	-0.0025	0.82
Tonhöhe	-0.0025	0.82
Tondauer	0.0074	0.49
Begl.-Tonhöhe	-0.0062	0.57

Tabelle 8.2: Auswertung über alle Instrumente und alle Plackett-Burman-Designs für die Instrumentenerkennung für HD1 mit optimiertem Hörgerät und *Random Forest*. Die Fehlerrate ist die Zielvariable (**dick** = signifikant zum 10%-Level).

Beim Vergleich der Ergebnisse darf allerdings nicht vergessen werden, dass die Fehlklassifikationsrate der Instrumentenerkennung nur ein Teilmaß ist. Das gefundene Optimum entspricht nicht unbedingt dem Optimum bezüglich des Gesamtmaßes, das im folgenden Abschnitt diskutiert wird. Außerdem bezieht sich das oben genannte Ergebnis lediglich auf die Optimierungsmenge Ω_{opt} und müsste eigentlich noch bezüglich einer unabhängigen Evaluierungsmenge Ω_{eval} verifiziert werden, auf die jedoch an dieser Stelle verzichtet

wird. Der ausgeführte Optimierungslauf ist trotz der Vereinfachungen und der Parallelisierung immer noch sehr zeitintensiv und benötigt über elf Tage. In Abschnitt 8.3.2 wird diskutiert, welche Möglichkeiten für eine effizientere Optimierung in Frage kommen.

8.3 Vervollständigung und Erweiterung der Optimierung

In diesem Abschnitt wird diskutiert, welche zusätzlichen Herausforderungen durch die vollständige Optimierung entstehen, bei der, neben der Instrumentenerkennung, auch die Fehlerraten der Einsatzzeiterkennung und der Tonhöhenschätzung verwendet werden. Dies beinhaltet zum einen die Bildung eines Gesamtmaßes aus den Kostenmaßen der drei Verfahren (Abschnitt 8.3.1), aber auch notwendige Schritte zur Rechenzeitreduktion (Abschnitt 8.3.2).

8.3.1 Kombination der Fehlermaße

Wenn nicht nur die Instrumentenerkennung, sondern auch die Einsatzzeiterkennung und die Tonhöhenerkennung berücksichtigt werden sollen, müssen die Maße (bzw. die Kosten) in geeigneter Weise verknüpft werden. Dies kann entweder durch eine Aggregation zu einem Gesamtmaß (bzw. einer Wünschbarkeitsfunktion) oder durch eine mehrkriterielle Optimierung gelöst werden. Da nicht davon auszugehen ist, dass die drei Fehlermaße gegenläufig sind, ist die folgende Aggregation eine intuitive Variante:

$$c_{\mathcal{A}}^{\text{plus}}(\boldsymbol{\theta}, \Omega_{\text{opt}}) = w_1 \cdot c_{\mathcal{A}}^{\text{onset}}(\boldsymbol{\theta}, \Omega_{\text{opt}}) + w_2 \cdot c_{\mathcal{A}}^{\text{pitch}}(\boldsymbol{\theta}, \Omega_{\text{opt}}) + w_3 \cdot c_{\mathcal{A}}^{\text{inst}}(\boldsymbol{\theta}, \Omega_{\text{opt}}), \quad \text{mit } w_1 + w_2 + w_3 = 1. \tag{8.6}$$

Dabei sind $c_{\mathcal{A}}^{\text{plus}}(\cdot)$ die Gesamtkosten und $c_{\mathcal{A}}^{\text{onset}}(\cdot)$, $c_{\mathcal{A}}^{\text{pitch}}(\cdot)$ und $c_{\mathcal{A}}^{\text{inst}}(\cdot)$ die Einzelkosten der Einsatzzeiterkennung, der Tonhöhenerkennung und der Instrumentenerkennung für einen Algorithmus $\mathcal{A}$ mit den Parametereinstellungen $\boldsymbol{\theta}$ gegeben eines Optimierungsdatensatzes (Musikstücke) Ω_{opt}. Für die Berechnung der Einzelkosten kommt neben der direkten Übernahme der Fehlklassifikationsraten als Alternative auch ein Vergleich zu den Fehlerraten ohne Hörschädigung in Betracht. Ein Nachteil von Definition 8.6 ist, dass sie keine Wechselwirkungen der Einzelmaße berücksichtigt, obwohl eine gewisse Abhängigkeit gegeben ist. Wenn beispielsweise kein einziger Toneinsatz richtig erkannt werden kann, ist es fraglich, ob eine Verbesserung der Tonhöhen- und Instrumentenerkennung genauso viel bringt, wie in dem Szenario, in dem alle Einsätze richtig erkannt werden.

Als Alternative kommt daher auch die folgende Variante in Betracht:

$$c_{\mathcal{A}}^{mult}(\boldsymbol{\theta}, \Omega_{\text{opt}}) = 1 - [1 - c_{\mathcal{A}}^{onset}(\boldsymbol{\theta}, \Omega_{\text{opt}})]^{w_1} \cdot [1 - c_{\mathcal{A}}^{pitch}(\boldsymbol{\theta}, \Omega_{\text{opt}})]^{w_2} \cdot [1 - c_{\mathcal{A}}^{inst}(\boldsymbol{\theta}, \Omega_{\text{opt}})]^{w_3}, \quad \text{mit } w_1 + w_2 + w_3 = 1. \tag{8.7}$$

Hier sind allerdings die Gesamtkosten $c_{\mathcal{A}}^{mult}(\boldsymbol{\theta}, \Omega_{\text{opt}})$ sehr stark von den höchsten Einzelkosten abhängig. Beispielsweise gilt $c_{\mathcal{A}}^{\text{inst}}(\boldsymbol{\theta}, \Omega_{\text{opt}}) = 1 \implies c_{\mathcal{A}}^{mult}(\boldsymbol{\theta}, \Omega_{\text{opt}}) = 1$, unabhängig davon, wie hoch die Einzelkosten der Einsatzzeit- und der Tonhöhenerkennung sind. Als guter Kompromiss erscheint daher eine Kombination aus $c_{\mathcal{A}}^{plus}(\cdot)$ und $c_{\mathcal{A}}^{mult}(\cdot)$ sinnvoll. Die optimale Parametrisierung des Gesamtmaßes kann jedoch nur durch einen Hörversuch ermittelt werden.

8.3.2 Zeitreduktion der Optimierung

Eine Zeitreduktion der Optimierung kann entweder durch eine Parallelisierung oder durch eine Datenreduktion erzielt werden. In Bischl u. a. (2014) haben wir eine MBO-Variante (MOI-MBO) entwickelt, bei der in jeder Iteration mehrere Punkte vorgeschlagen werden, die dann parallel ausgewertet werden können. Die Grundidee ist die Definition eines mehrkriteriellen Infill-Kriteriums, das, neben der erwarteten Verbesserung, als weiteres Kriterium die Distanz (die maximiert werden soll) zum nächsten anderen auszuwertenden Punkt berücksichtigt. Dieses Infill-Kriterium wird mit Hilfe eines mehrkriteriellen Evolutionären Algorithmus bezüglich des Hypervolumens der Paretofront optimiert. Alle Punkte der Endpopulation werden dann für die Auswertung vorgeschlagen. Die Idee dahinter ist, dass diese Punkte sowohl eine hohe erwartete Verbesserung aufweisen, aber gleichzeitig auch weit voneinander entfernt sind, so dass deren Auswertung einen maximalen Informationsgewinn mit sich bringt. Allerdings ist die Optimierung auch intern stark parallelisierbar, da jedes Musikstück unabhängig transformiert werden kann. Da eine innere Optimierung prinzipiell zu bevorzugen ist, ist es fraglich, ob MOI-MBO für das konkrete Optimierungsproblem in Frage kommt.

Die zweite Möglichkeit zur Rechenzeitreduzierung ist die Reduzierung der Daten. In Bauer u. a. (2016) haben wir ein Verfahren (FMBO) eingeführt, das in jeder Iteration zunächst einen neuen Punkt $\boldsymbol{\theta}_{neu}$ nur auf einer Teilmenge $\Omega_{\text{opt}}^{pretest} \subset \Omega_{\text{opt}}$ der Instanzen evaluiert. Basierend auf den Ergebnissen der Einzelkostenfunktionen $c_{\mathcal{A}}^{\omega}(\omega, \boldsymbol{\theta}_{neu})$ mit $\omega \in \Omega_{\text{opt}}^{pretest}$ wird anschließend entschieden, ob dieser Punkt als aussichtsreich eingestuft

wird. Nur in diesem Fall wird er anschließend auch auf den restlichen Punkten von Ω_{opt} ausgewertet. Andernfalls wird lediglich eine Schätzung von $c_{\mathcal{A}}(\boldsymbol{\theta}_{neu}, \Omega_{\text{opt}})$ für die Aktualisierung des Surrogatmodells verwendet. In dem Artikel haben wir das Verfahren für die Optimierung des Algorithmus für die Einsatzzeiterkennung entwickelt, er ist aber prinzipiell für jedes instanzbasierte Optimierungsproblem adaptierbar und somit auch für die Hörgeräteoptimierung.

Zunächst müssen die Punkte des initialen Designs Θ_{init} vollständig auf allen Instanzen Ω_{opt} evaluiert werden, um eine ausreichend große Datengrundlage zu schaffen. Anschließend wird die Teilmenge $\Omega_{\text{opt}}^{pretest}$ so ausgewählt, dass die enthaltenen Punkte möglichst repräsentativ sind. In dem Artikel wird dafür ein Cluster-Verfahren verwendet, das die Instanzen gemäß der Ähnlichkeit ihrer bisherigen Kosten $c_{\mathcal{A}}^{\omega}(\boldsymbol{\theta}, \omega)$ mit $\boldsymbol{\theta} \in \Theta_{\text{init}}$ und $\omega \in \Omega_{\text{opt}}$ zu Gruppen zusammenfasst. Aus jeder Gruppe wird anschließend zufällig ein Punkt ausgewählt, und die Menge dieser ausgewählten Punkte ergibt die Teilmenge $\Omega_{\text{opt}}^{pretest}$. Mit diesen Punkten und allen bisherigen Auswertungen kann dann ein lineares Modell (oder auch ein anderes Regressionsmodell) angepasst werden, das für einen neuen Punkt $\boldsymbol{\theta}_{neu}$, gegeben aller Einzelkosten $c_{\mathcal{A}}^{\omega}(\boldsymbol{\theta}, \omega)$ der Instanzen $\omega \in \Omega_{\text{opt}}^{pretest}$, die Gesamtkosten der Optimierungsmenge $c_{\mathcal{A}}(\boldsymbol{\theta}, \Omega_{\text{opt}})$ schätzt. Zudem wird auch eine Unsicherheit mitgeschätzt, so dass berechnet werden kann, wie wahrscheinlich der neue Punkt teurer als das bisherige Minimum $c_{\mathcal{A}}^{\min}$ ist:

$$P\left(c_{\mathcal{A}}(\boldsymbol{\theta}, \Omega_{\text{opt}}) > c_{\mathcal{A}}^{\min} \mid c_{\mathcal{A}}^{\omega}(\boldsymbol{\theta}, \omega) \forall \omega \in \Omega_{\text{opt}}^{\text{pretest}}\right). \tag{8.8}$$

Wenn dieser Wert kleiner als ein kritischer Wert p ist, wird der Punkt $\boldsymbol{\theta}_{\text{neu}}$ als aussichtsreich eingestuft und vollständig auf allen Instanzen ausgewertet. In dem zitierten Artikel wurde $p = 0.99$ gesetzt.

Im Gegensatz zur Optimierung der Einsatzzeiterkennung, für die wir das Verfahren in dem Artikel getestet haben, ergibt sich für die Hörgeräteoptimierung ein wichtiger Unterschied. Während bei der Optimierung der Einsatzzeiterkennung die Kosten einer Instanz lediglich von der Parameterkonfiguration abhängen, basieren die Kosten bei der Hörgeräteoptimierung auf den Fehlklassifikationsraten interner Klassifikationsmodelle für die Einsatzzeit-, Tonhöhen- und Instrumentenerkennung. Diese Modelle müssen in jeder Optimierungsiteration erneut auf die neu transformierten Daten angepasst werden, was im Falle der Einsatzzeiterkennung einer weiteren inneren Parameteroptimierung entspricht. Die Klassifikationsgüten sind jedoch nicht nur von der Parameterkonfiguration des Hörgeräts abhängig (durch die die Trennbarkeit der Daten beeinflusst wird), sondern auch

von der Trainingsmenge, insbesondere von deren Größe. Ein kleinerer Datensatz erhöht den zu erwartenden Fehler und zudem auch dessen Varianz (Brain u. a., 1999). Daher ist die Einzelkostenfunktion einer Instanz $\omega \in \Omega$ auch von der verwendeten Trainingsmenge Ω_{train} abhängig und wird somit genauer durch $c^{\omega}_{\mathcal{A}} : \Omega \times \Theta \times \Omega^{|\Omega_{\text{train}}|} \rightarrow \mathbb{R}$ ausgedrückt.[5] Falls nun die Bedingung $\Omega_{\text{train}} \subseteq \Omega^{\text{pretest}}_{\text{opt}}$ nicht gilt, kann $c^{\omega}_{\mathcal{A}}(\omega, \boldsymbol{\theta}, \Omega_{\text{train}})$ nicht berechnet werden und muss stattdessen durch $c^{\omega}_{\mathcal{A}}(\omega, \boldsymbol{\theta}, \Omega^{\text{pretest}}_{\text{train}})$ mit $\Omega^{pretest}_{\text{train}} = \Omega_{\text{train}} \cap \Omega^{\text{pretest}}_{\text{opt}}$ approximiert werden. Es wird also eine kleinere Trainingsmenge verwendet, wodurch der Erwartungswert und die Varianz der Einzelkosten steigen. Daher ist auch für die Schätzung der wahren Gesamtkosten $c_{\mathcal{A}}(\boldsymbol{\theta}, \Omega_{\text{opt}})$ – die natürlich weiterhin auf den vollständigen Trainingsmengen basieren – ein höherer Fehler zu erwarten. Die Anwendbarkeit von FMBO für die Hörgeräteoptimierung hängt somit davon ab, wie gut diese Kosten auf Basis der Teilmenge $\Omega^{\text{pretest}}_{\text{opt}}$ und der damit verbundenen kleineren Trainingsmengen $\Omega^{pretest}_{\text{train}}$ geschätzt werden können.

Eine weitere Möglichkeit die Anzahl der Instanzen zu reduzieren, ist eine Reduktion des Versuchsplans. Beispielsweise können Faktoren, die nie signifikante Einflüsse zeigen, aus dem Versuchsdesign entfernt werden. Unter Berücksichtigung aller in dieser Arbeit ausgewerteten Experimente, trifft dies allerdings auf keinen der acht untersuchten Faktoren zu. Statt die Instanzen (Musikstücke) zu reduzieren, kann auch ihre Länge reduziert werden. Die Instanzlänge von 30 Melodietönen stellt einen guten Kompromiss in Bezug auf Rechenzeit und Modellierungsgüte dar, ist aber dennoch relativ willkürlich festgelegt.

Für die Ermittlung einer angemessenen Datengröße bietet sich die Untersuchung der Lernkurve an, die den Trainingsfehler ϵ_{train} und den Testfehler ϵ_{test} in Abhängigkeit zur Größe l der Trainingsmenge setzt. Der Trainingsfehler eines Klassifikationsmodells ist die Fehlklassifikationsrate bezüglich der Trainingsmenge und dementsprechend der Testfehler die Fehlerrate bezüglich der Testmenge. Auf Grund einer Überanpassung bei kleineren Trainingsmengen, steigt ϵ_{train}, je größer die Trainingsmenge ist, wohingegen ϵ_{test} sinkt. Für große l nähern sich beide Fehler asymptotisch dem tatsächlichen Fehler a an:

$$\epsilon_{\text{test}} = a + \frac{b}{l^{\alpha}} \quad \text{und} \quad \epsilon_{\text{train}} = a - \frac{c}{l^{\beta}}, \tag{8.9}$$

wobei b, c, α und β positive Werte sind, die vom Klassifikationsverfahren und Datensatz abhängen und den Lernverlauf beschreiben. Durch einen Vergleich von ϵ_{train} und ϵ_{test} können sowohl der wahre Fehler a als auch die durch eine Vergrößerung der Trainingsmenge erwartete Verbesserung des Klassifikationsmodells geschätzt werden (Cortes u. a., 1994).

[5] Der Einfachheit halber wird angenommen, dass das Klassifikationsmodell deterministisch erstellt wird.

Eine weitere Möglichkeit besteht darin, die Signifikanz der Klassifikationsgüte bei einer gegebenen Datengröße durch einen statistischen Test zu ermitteln. In Mukherjee u. a. (2003) wird dafür ein Permutationstest verwendet, der die Güte des Klassifikationsmodells mit der Güte eines randomisierten Klassifikationsmodells vergleicht, das auf den gleichen Daten, aber mit zufällig permutierten Klassenlabeln trainiert ist.

9 Zusammenfassung und Ausblick

Ein automatisches Bewertungsmaß, das unabhängig von Hörversuchen die Verständlichkeit von Sprache oder Musik misst, ist wichtig für die Optimierung von Hörgeräten. Ein natürlicher Ansatz dafür ist die Verwendung von Simulationsmodellen der auditorischen Peripherie (Ohrmodell), wodurch die im menschlichen Ohr stattfindende Transformation von eintreffenden Schallwellen in Hörnervenimpulse simuliert werden kann. Während diese Modelle schon umfangreich validiert sind, ist der darauf im Gehirn anschließende Interpretationsvorgang, der aus den Hörnervenimpulsen die menschliche Wahrnehmung von Geräuschen, Sprache oder Musik bildet, bei weitem noch nicht so ausreichend erforscht, um ein vollständiges Modell der Hörwahrnehmung erstellen zu können.

In dieser Arbeit wird ein Konzept zur automatischen Bewertung von Hörgerätealgorithmen für Musik präsentiert, das, basierend auf der Ohrmodellausgabe, die daran anschließende Musikwahrnehmung durch statistische Klassifikationsverfahren schätzt. Wenn die Ohrmodellausgabe durch die Simulation einer typischen Hörschädigung (beispielsweise der Ausfall von inneren oder äußeren Haarzellen) beeinträchtigt ist, werden, analog zur menschlichen Musikwahrnehmung, auch die Klassifikationsmodelle ungenauer. Mit Hörgeräten kann dieser Effekt kompensiert werden, z.B. indem bestimmte Frequenzbereiche verstärkt werden oder die Komplexität so reduziert wird, dass die verbliebenen Musikmerkmale weniger verrauscht sind und entsprechend besser wahrgenommen bzw. klassifiziert werden. Durch eine Verknüpfung mit der Bewertungsfunktion kann somit für einen vorgeschlagenen Hörgerätealgorithmus evaluiert werden, wie gut damit bei einer gegebenen Hörbeeinträchtigung Musik differenziert werden kann.

Um die Problemstellung besser definieren zu können, wird das Anwendungsgebiet auf klassische Kammermusikstücke mit einer klar definierten Melodiestimme (und beliebig vielen Begleitstimmen) beschränkt. Das Ziel besteht darin, die musikalischen Eigenschaften der Melodiestimme so gut wie möglich zu identifizieren. Konkret werden drei wesentliche musikalische Eigenschaften untersucht, für deren Schätzung vorhandene

klassifikationsbasierte Verfahren untersucht und weiterentwickelt werden: (1) Die Toneinsatzzeiterkennung, (2) die Tonhöhenerkennung und (3) die Instrumentenerkennung. Die Fehlerraten dieser drei Klassifikationsverfahren definieren dann, wie gut ein Hörgerätealgorithmus – mit einer bestimmten Parametereinstellung – für eine spezifische Hörschädigung die Wahrnehmung verbessert. Es werden verschiedene Möglichkeiten für die Aggregation der drei Einzelfehlerraten vorgeschlagen, deren exakte Parametrisierung jedoch noch durch Hörversuche zu spezifizieren ist. In der Zukunft könnte das Konzept durch die Schätzung weiterer Musikeigenschaften erweitert werden, beispielsweise durch eine Pausenerkennung, eine Melodieerkennung, welche die zeitliche Abfolge von Tönen berücksichtigt oder durch eine Lautheitsschätzung, die jedoch mathematisch schwierig zu spezifizieren ist. Eine weitere Möglichkeit ist die Berücksichtigung der Begleitung, die in der jetzigen Form lediglich als störendes Rauschen betrachtet wird. Eine Idee ist, auch deren Merkmale zu schätzen, aber die entsprechenden Fehlermaße mit einer geringeren Gewichtung in das Gesamtmaß einfließen zu lassen. Je nach Hörgerätealgorithmus kann es zudem auch notwendig sein, weitere technische Maße mit einfließen zu lassen, um den Erhalt von bestimmten musikalischen Strukturen zu gewährleisten. Für den in dieser Arbeit untersuchten Algorithmus ist dies jedoch nicht nötig.

Für die Schätzung der drei in dieser Arbeit untersuchten Eigenschaften existieren umfangreiche Forschungsarbeiten, die jedoch üblicherweise nicht die Ohrmodellausgabe sondern die akustische Wellenform als Grundlage nutzen. In dieser Arbeit wird gezeigt, dass viele dieser Standardverfahren durch leichte Modifikationen auf die Ohrmodell-basierte Schätzung übertragen werden können. Dafür werden in der Regel die zu extrahierenden Merkmale für jeden Kanal des Ohrmodells unabhängig berechnet. Eine Herausforderung besteht darin, die Merkmale oder Schätzungen sinnvoll zu kombinieren, auch in Hinsicht auf Rechenzeitaspekte.

Für die Einsatzzeiterkennung mit Ohrmodell werden verschiedene Möglichkeiten für die Aggregation mehrerer Kanäle (Merkmale) untersucht. Das Verfahren, das zunächst die Schätzungen für jeden Kanal vollständig unabhängig vollzieht und diese erst zum Schluss kombiniert, schneidet am besten ab. Bei diesem Verfahren wird jeder Zeitpunkt, der mindestens einmal als Toneinsatz identifiziert ist, als Kandidat betrachtet, und anschließend wird er, abhängig von der Anzahl von weiteren Kandidaten in der zeitlichen Umgebung, bestätigt oder verworfen. Durch dieses Konzept wird die Einsatzzeiterkennung effizient in ein klassisches Klassifikationsproblem überführt. Eine Erweiterung durch zusätzliche Merkmale ist daher nun problemlos möglich und erscheint sinnvoll.

Das Verfahren kann auch für die Einsatzzeiterkennung ohne Ohrmodell verwendet werden, um die Einzelschätzungen mehrerer Merkmale – z.B. *Spectral Flux* und Phasenverschiebung – zu aggregieren. Durch die Verwendung der 18 in Bauer (2016) untersuchten Merkmale würde das Verfahren prinzipiell sehr ähnlich zu der dort entwickelten multivariaten Klassifikationsmethode. Im Unterschied zu dieser wird jedoch nicht jeder Zeitpunkt (Fensternummer) als Beobachtung betrachtet, sondern nur diejenigen Zeitpunkte, die von mindestens einem Merkmal als Toneinsatz identifiziert sind. Dadurch ergeben sich einige Vorteile:

- Es kann viel Rechenzeit eingespart werden, da die Anzahl der Beobachtungen geringer ist. Diese Einsparung kann auch dafür verwendet werden, die Klassifikationsgüte durch die Hinzunahme weiterer Daten oder Merkmale zu verbessern.
- Zudem ist die Klassenverteilung weniger unausgeglichen, denn der Anteil an positiven Beobachtungen (Zeitpunkte, an denen Töne einsetzen) erhöht sich durch die starke Reduktion von negativen Beobachtungen.
- Die Einzelschätzungen dürfen asynchron sein (d.h. unterschiedliche Fenstergrößen M, Sprungweiten h und Verschiebungskorrekturen τ sind erlaubt).

Dem gegenüber stehen zwei potentielle Nachteile:

- Ein Toneinsatz kann nur dann erkannt werden, wenn er von mindestens einer Einzelschätzung als Kandidat identifiziert ist. Dagegen könnte das in Bauer (2016) entwickelte Klassifikationsverfahren den Toneinsatz möglicherweise dennoch finden.
- Eine leichte Latenz (Verschiebung) einer Toneinsatzerkennung in allen Einzelschätzungen kann vom Klassifikationsverfahren nicht mehr korrigiert werden.

Die tatsächlichen Auswirkungen dieser Überlegungen sollten durch eine experimentelle Vergleichsstudie verifiziert werden.

Auch die Klangfarbenmerkmale, die in Kapitel 5.3 für die Instrumentenerkennung definiert sind, erscheinen für die Einsatzzeiterkennung eine sinnvolle Erweiterung, denn die je nach Musikinstrument unterschiedlichen Ausprägungen von Toneinsätzen sollten vom Klassifikationsmodell berücksichtigt werden (Böck und Widmer, 2013). Vor allem für die predominante Einsatzzeiterkennung ist einer Erweiterung der Merkmalsmenge wichtig, um besser Melodie- und Begleittoneinsätze trennen zu können. Eine weitere Möglichkeit zur Verbesserung der Einsatzzeiterkennung ist eine Quantisierung der Schätzungen (Bauer,

2016). Diese berücksichtigt die regelmäßige Struktur von rhythmischer Musik und lässt nur zu bestimmten Zeitpunkten die Schätzung eines Toneinsatzes zu.

Im Gegensatz zur Tonhöhen- und Instrumentenerkennung, bei denen die Klassifikation mit Ohrmodellmerkmalen sehr gute Ergebnisse liefert, ist die automatische Einsatzzeiterkennung recht fehleranfällig. Dies gilt aber auch für das Standardverfahren ohne Ohrmodell, das selbst bei den Experimenten mit monophoner Musik einen F-Wert von lediglich 0.86 erreicht, während bei einem menschlichen Hörer ein F-Wert nahe 1 erwarten werden kann (alle Töne haben eine Mindestlänge von 0.1 Sekunden und keine Töne sind unnatürlich leise). Aus diesen Gründen erscheint es fraglich, ob die Leistung der automatischen Einsatzzeiterkennung ausreichend ist, um hörschädigungsbedingte Wahrnehmungsschwächen aufzudecken.

Die Fehlerraten sind allerdings für die drei untersuchten Musikinstrumente sehr unterschiedlich. Während die Einsätze der Trompete fast perfekt erkannt werden, ist vor allem die Erkennung von Celloeinsätzen problematisch. Eine Möglichkeit ist es daher, für das Bewertungsmaß der Einsatzzeiterkennung nur Instrumente zu berücksichtigen, mit denen akzeptable Ergebnisse erzielt werden. Aktuelle Studien mit Schwerhörigen deuten allerdings auch darauf hin, dass die Verbesserung von Tonhöhen- und Instrumentenerkennung sowieso wichtiger als die Verbesserung der Toneinsatzerkennung ist. In mehreren Hörversuchen mit Hörgeräte- und Cochlea-Implantat-Trägern hat es sich herausgestellt, dass diese im Vergleich zu Normalhörenden deutlich schlechter bei der Unterscheidung von Instrumenten, Tonhöhen und Melodien abschneiden, aber es bei der Erkennung von Rhythmen keine signifikanten Unterschiede gibt (McDermott, 2004; Gfeller u. a., 2006; Looi u. a., 2008).

Bei der predominanten Tonhöhenerkennung schneidet das entwickelte klassifikationsbasierte Verfahren unter Verwendung von spektralen Merkmalen am besten ab. Bei diesem Verfahren wird zunächst der Kanal klassifiziert, dessen *Best Frequenz* der Grundfrequenz am nächsten ist, das heißt die Klassifikation schränkt die Suche auf einen bestimmten Frequenzbereich ein, der durch die Bandweite des Kanals beschränkt wird. Anschließend wird innerhalb dieser Bandweite die Frequenz als Grundfrequenz geschätzt, deren spektrale Amplitude in dem Kanal maximal ist. Unter Verwendung einer linearen SVM für die Kanalklassifikation, wird in den Experimenten eine Fehlerrate von 7% erreicht (bei einer Fehlertoleranz von einem halben Halbton). Des Weiteren wird gezeigt, dass die hohe Anzahl an Merkmalen, die das Verfahren verwendet ($41 \cdot 29 = 1189$ Merkmale) durch eine auf Merkmalsgruppen erweiterte Rückwärtsselektion stark reduziert werden kann, ohne

dass das Klassifikationsmodell an Genauigkeit verliert. Bei der Kanal-basierten Gruppierung wird dabei die Anzahl auf 667 und bei der Merkmalstyp-basierten Gruppierung sogar auf 369 Merkmale reduziert. Als Erweiterung drängt sich eine Kombination der beiden Selektionsvarianten auf, um die Anzahl der Merkmale noch weiter zu reduzieren. Die Merkmale, welche über alle Kanäle die maximale Intensität der potentiellen Obertöne beschreiben $P_{pl}^{max}[k]$, werden in keiner Variante von der Merkmalsselektion ausgewählt und scheinen daher überflüssig zu sein. Zur weiteren Verbesserung des Verfahrens zur Tonhöhenerkennung sollte die Bestimmung der Frequenz mit maximaler Amplitude durch die Verwendung eines Interpolationsverfahrens (z.B. Quinn, 1994) präzisiert werden. Dazu kann auch zusätzliche Information aus den beiden umliegenden Kanälen oder aus der Autokorrelationsfunktion beitragen. Überhaupt bietet es sich auch für die Klassifikation an, sowohl Merkmale aus dem Frequenzbereich als auch aus dem Zeitbereich (Autokorrelation) zu verwenden, anstatt sich nur auf einen Typ von Merkmalen zu beschränken. Schließlich wird auch bei der menschlichen Tonhöhenwahrnehmung davon ausgegangen, dass beide Bereiche kombiniert werden (Langner, 1981).

Die Instrumentenerkennung ist ein klassisches Klassifikationsproblem, bei dem die Merkmale, die aus den einzelnen Kanälen extrahiert werden, problemlos zu einem großen Merkmalssatz zusammengefasst werden können. Die verwendeten Merkmale entsprechen dabei Standardmerkmalen, die aber, anstatt aus der Schallwellenform, aus den Kanalausgaben extrahiert werden. Mit Hilfe dieser Merkmale können bei der predominanten Instrumentenerkennung die drei untersuchten Musikinstrumente (Cello, Klarinette und Trompete) mit einer Fehlklassifikationsrate von 1.1% fast perfekt getrennt werden, sowohl durch eine lineare SVM als auch durch einen *Random Forest* (mit Merkmalsselektion). Bei Verwendung der Standardmerkmale (ohne Ohrmodell) wird dagegen als beste Fehlerrate nur 3.5% erreicht. Besonders relevant sind Merkmale der höheren Kanäle, die bei den üblichen Hörschädigungen am stärksten beeinträchtigt sind. Für den *Random Forest* sind 12 Kanäle, also $12 \cdot 21 = 252$ Merkmale ausreichend, um die niedrigste Fehlklassifikationsrate zu erhalten. Der Vorteil der auditorischen Merkmale gegenüber den Standardmerkmalen ist vermutlich einer geringeren Verrauschung der Merkmale durch gleichzeitig spielende Instrumente zuzuschreiben. Durch die spektrale Aufteilung eines Signals in die Kanäle des Ohrmodells, können die meisten extrahierten Merkmale vorwiegend einer Quelle zugeordnet werden. Diese These wird auch dadurch untermauert, dass nur bei der predominanten Instrumentenerkennung die auditorischen Merkmale besser abschneiden, während bei der monophonen Variante, bei der keine Vermischung verschiedener Quellen auftritt, die Standardmerkmale genauso gute Ergebnisse erzielen.

Bei diesen Ergebnissen muss berücksichtigt werden, dass die drei untersuchten Instrumente sehr verschieden sind. In Zukunft sollte daher auch ein umfangreicherer Datensatz untersucht werden. Bei der in Kapitel 5.4 beschriebenen Vorstudie wurde bei fünf Instrumenten (Flöte, Klarinette, Oboe, Trompete und Violine) nur eine beste Fehlklassifikationsrate von 12% erreicht (siehe Tabelle 5.4). Allerdings wurden dort deutlich weniger Beobachtungen verwendet (765 Beobachtungen in der Trainingsmenge gegenüber 2160 bei den neueren Experimenten) und es wurde nicht sichergestellt, dass die vordefinierte Melodiestimme auch tatsächlich dominant ist.

In der Literatur finden sich noch eine Reihe weiterer Merkmale für die Instrumentenklassifikation, die in dieser Arbeit nicht berücksichtigt sind. Für eine Erweiterung bieten sich vor allem Merkmale an, die den Anschlag eines Tons beschreiben. Solche Merkmale sind beispielsweise in Eronen und Klapuri (2000) definiert. Auch bei der menschlichen Wahrnehmung spielt der Anschlag eine wichtige Rolle für die Unterscheidung von Musikinstrumenten (Roederer und Mayer, 1999). Weiterhin bieten sich auch die in Martin (1999) vorgeschlagenen Merkmale an. Diese beschreiben die Verteilung der Feuerraten auf die Kanäle, wohingegen die meisten in dieser Arbeit verwendeten Merkmale die Frequenzverteilung der Feuerraten innerhalb der Kanäle beschreiben.

Sowohl für die Instrumentenerkennung als auch für die Tonhöhenerkennung sind viele Merkmale eher willkürlich gewählt. Beispielsweise ist nicht ersichtlich, warum *Spectral Rolloff 85* als Merkmal verwendet wird und nicht *Spectral Rolloff 90* (siehe Gleichung 5.5). Da viele von diesen Merkmalen einfach parametrisierbar sind – wie der Wert R beim Merkmal *Spectral Rolloff* –, kommt der Einsatz einer Optimierung (z.B. MBO) stark in Betracht, um einen verbesserten Merkmalssatz zu erhalten.

Alle Klassifikationsverfahren werden in dieser Arbeit auch für drei auditorische Modelle mit unterschiedlichen Hörschädigungen (*Hearing Dummies*) getestet. Für all diese Modelle steigen die Fehlerraten in plausiblen Stärken, die abhängig von den Hörschädigungen sind. Eine exakte Einordnung dieser Ergebnisse ist allerdings nur mit zukünftigen Hörversuchen möglich. Diese sollten auch dafür verwendet werden, die Klassifikationsmodelle durch eine Hyperparameteroptimierung an die Realität anzupassen.

Für die Evaluierung der Verfahren in dieser Arbeit wird ein statistischer Versuchsplan, dem ein Plackett-Burman-Design zu Grunde liegt, verwendet. Dadurch werden die Daten (Musikstücke) in einer strukturierten Form auf eine akzeptable Größe reduziert. Zudem können auch Aussagen darüber getroffen werden, für welche Art von Musik ein Verfahren besonders gut oder schlecht funktioniert. Durch den Versuchsplan werden

acht musikalische Einflussgrößen berücksichtigt, von denen sieben durch einen Plackett-Burman-Plan (je zwei Niveaus) mit zwölf Experimenten modelliert werden. Die achte Einflussgröße ist das Melodieinstrument, für das drei Niveaus untersucht werden. Die Kombination geschieht dabei vollfaktoriell, so dass der Plan insgesamt aus $3 \cdot 12 = 36$ Versuchen ($\widehat{=}$ Musikstücken) besteht. Ein Problem bei der Verwendung des Plackett-Burman-Designs ist, dass Wechselwirkungen zwischen den Einflussgrößen nicht modelliert werden. Allerdings würde ein vollfaktorieller Plan $3 \cdot 2^7 = 384$ Versuche beinhalten, was aus Rechenzeitgründen nicht durchführbar ist.

Mit Hilfe des Versuchsplans kommen neben vielen erwarteten Ergebnissen, z.B. die größeren Fehlerraten bei einer Streicherbegleitung auf Grund der klanglichen Nähe zum Cello, auch einige unerwartete Ergebnisse heraus. Beispielsweise sind höhere Tonhöhen und kürzere Töne vorteilhaft für die predominante Einsatzzeiterkennung, wohingegen tiefere Tonhöhen die Ergebnisse der Instrumentenerkennung verbessern. Für zukünftige Studien könnten noch weitere Einflussgrößen, wie die Anzahl der Begleitinstrumente oder die Varianz der Tonlängen untersucht werden. Demgegenüber könnte man auch Einflussgrößen, die in keinem Experiment einen signifikanten Einfluss haben, weglassen und somit den Versuchsaufbau vereinfachen.

In Kapitel 8 wird die praktische Anwendbarkeit des Verfahrens in einer leicht vereinfachten Form, die aus Rechenzeitgründen lediglich die Ergebnisse der Instrumentenerkennung berücksichtigt, für die Optimierung eines Hörgerätealgorithmus getestet. Dabei wird MBO verwendet, um das Hörgerät optimal an eine starke Hörschädigung (*Hearing Dummy* 1) anzupassen. Durch das optimierte Hörgerät wird die Fehlklassifikationsrate stark reduziert, und auch eine vergleichende Experteneinstellung wird geschlagen (27% ohne Hörgerät, 19% mit Hörgerät und Experteneinstellung, 14% mit optimiertem Hörgerät). Wie die Auswertung des Versuchsplans zeigt, wird vor allem die Klassifikationsgüte für Musikstücke mit Streicherbegleitung verbessert. Für die Zukunft ist die größte Herausforderungen der Hörgeräteoptimierung die Reduktion der Rechenzeit, wozu bereits in Kapitel 8.3.2 ein Ausblick beschrieben ist.

Ein weiterer Ausblick ist die Integration der verschiedenen Optimierungsebenen in ein gemeinsames Modell, das eine effizientere simultane Optimierung ermöglicht. Die verschiedenen Optimierungsstufen sind in Abbildung 9.1 skizziert. Neben der unmittelbaren Optimierung des Hörgeräts mit den Parametern zur Einsatzzeiterkennung, Komplexitätsreduktion und Dynamikanpassung, kommt auch eine Optimierung der Klassifikationsmodelle und ihrer Merkmale in Betracht, um die Bewertungsmodelle besser an die transformierten

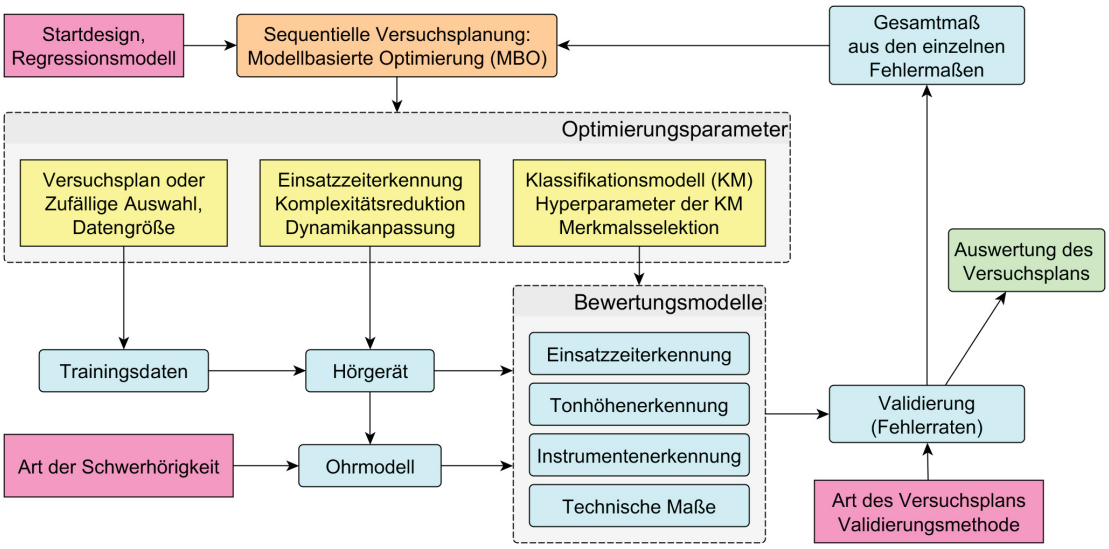

Abbildung 9.1: Optimierung auf mehreren Ebenen (Ausblick).

Daten anzupassen. Die Optimierung der Daten, insbesondere deren Umfang, ist essentiell für eine schnellere Optimierung. Eine Optimierung des Datenumfangs ist jedoch nur unter Berücksichtigung des Zeitbedarfs sinnvoll, da andernfalls eine größere Trainingsmenge immer bevorzugt würde. Eine Möglichkeit dafür ist es, zusätzlich auch den erwarteten Zeitbedarf einer Parametereinstellung zu schätzen und das *Expected Improvement* im Verhältnis zur Rechenzeit zu betrachten. Außer den Optimierungsparametern ist auch eine Optimierung der Hyperparameter von Interesse, denn der Hörgerätealgorithmus muss nicht nur einmal, sondern für jede Art der Schwerhörigkeit erneut optimiert werden. Dabei erscheint es sinnvoll, die bisherigen Ergebnisse auch für die Optimierung einer neuen Hörschädigung zu verwenden, beispielsweise durch eine Anpassung des initialen Designs. Eine weitere Möglichkeit besteht darin, die Art der Hörschädigung parametrisiert zu erfassen. Dadurch wäre die Schätzung eines gemeinsamen Surrogatmodells möglich. Für eine gemeinsame Optimierung der verschiedenen Ebenen ist es wichtig, die Abhängigkeitsstruktur der Parameter zu ermitteln. Unabhängige Parametersätze können dann durch eine separate Behandlung effizienter optimiert werden.

Literatur

Bauer, N., K. Friedrichs, D. Kirchhoff, J. Schiffner und C. Weihs (2014). „Tone Onset Detection Using an Auditory Model“. In: *Data Analysis, Machine Learning and Knowledge Discovery*. Hrsg. von Myra Spiliopoulou, Lars Schmidt-Thieme und Ruth Janning. Bd. Part VI. Hildesheim, Germany: Springer International Publishing, S. 315–324. DOI: 10.1007/978-3-319-01595-8_34.

Bauer, Nadja (2016). „Optimierung der Toneinsatzzeiterkennung“. Diss. Department of Statistics, TU Dortmund University.

Bauer, Nadja, Klaus Friedrichs, Bernd Bischl und Claus Weihs (2016). „Fast model based optimization of tone onset detection by instance sampling“. In: *Analysis of Large and Complex Data (to appear)*. Hrsg. von Hans A. Kestler Adalbert F.X. Wilhelm. Bremen, Germany: Springer International Publishing.

Bello, Juan Pablo, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies und Mark B Sandler (2005). „A tutorial on onset detection in music signals“. In: *IEEE Transactions on Speech and Audio Processing* 13.5, S. 1035–1047. DOI: 10.1109/TSA.2005.851998.

Bischl, Bernd, Simon Wessing, Nadja Bauer, Klaus Friedrichs und Claus Weihs (2014). „MOI-MBO: multiobjective infill for parallel model-based optimization“. In: *Learning and Intelligent Optimization*. Gainesville, FL, USA: Springer, S. 173–186. DOI: 10.1007/978-3-319-09584-4_17.

Bischl, Bernd, Michel Lang, Olaf Mersmann, Jörg Rahnenführer und Claus Weihs (2015). „BatchJobs and BatchExperiments: Abstraction Mechanisms for Using R in Batch Environments“. In: *Journal of Statistical Software* 64.11, S. 1–25. DOI: 10.18637/jss.v064.i11.

Bischl, Bernd, Michel Lang, Jakob Richter, Jakob Bossek, Leonard Judt, Tobias Kuehn, Erich Studerus und Lars Kotthoff (2016a). *mlr: Machine Learning in R*. R package version 2.5. URL: https://github.com/mlr-org/mlr.

Bischl, Bernd, Jakob Bossek, Daniel Horn und Michel Lang (2016b). *mlrMBO: Model-Based Optimization for mlr*. R package version 1.0. URL: https://github.com/berndbischl/mlrMBO.

Böck, Sebastian, Florian Krebs und Markus Schedl (2012). „Evaluating the Online Capabilities of Onset Detection Methods." In: *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, S. 49–54.

Böck, Sebastian und Gerhard Widmer (2013). „Maximum filter vibrato suppression for onset detection". In: *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13), Maynooth, Ireland*, S. 55–61.

Brain, Damien, G Webb, D Richards, G Beydoun, A Hoffmann und P Compton (1999). „On the effect of data set size on bias and variance in classification learning". In: *Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales*, S. 117–128.

Breiman, L. (2001). „Random Forests". In: *Machine Learning Journal* 45.1, S. 5–32. DOI: 10.1023/A:1010933404324.

Breiman, Leo (1996). „Bagging predictors". In: *Machine learning* 24.2, S. 123–140. DOI: 10.1007/BF00058655.

Cortes, Corinna, Lawrence D Jackel, Sara A Solla, Vladimir Vapnik und John S Denker (1994). „Learning curves: Asymptotic values and rate of convergence". In: *Advances in Neural Information Processing Systems*, S. 327–334.

Davis, S. und P. Mermelstein (1980). „Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4, S. 357–366. DOI: 10.1109/TASSP.1980.1163420.

De Cheveigné, Alain und Hideki Kawahara (2002). „YIN, a fundamental frequency estimator for speech and music". In: *The Journal of the Acoustical Society of America* 111.4, S. 1917–1930. DOI: 10.1121/1.1458024.

Dixon, Simon (2006). „Onset detection revisited". In: *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06)*, S. 133–137.

Duan, Zhiyao, Bryan Pardo und Changshui Zhang (2010). „Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.8, S. 2121–2133. DOI: 10.1109/TASL.2010.2042119.

Eerola, Tuomas und Rafael Ferrer (2008). „Instrument library (MUMS) revised". In: *Music Perception: An Interdisciplinary Journal* 25.3, S. 253–255.

Emiroglu, Suzan und Birger Kollmeier (2008). „Timbre discrimination in normal-hearing and hearing-impaired listeners under different noise conditions“. In: *Brain research* 1220, S. 199–207. DOI: 10.1016/j.brainres.2007.08.067.

Eronen, A. und A. Klapuri (2000). „Musical instrument recognition using cepstral coefficients and temporal features“. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, S. II–753–II–756. DOI: 10.1109/ICASSP.2000.859069.

Eronen, Antti (2001). „Automatic musical instrument recognition“. In: *Mémoire de DEA, Tempere University of Technology*, S. 178.

Essid, S., G. Richard und B. David (2006). „Hierarchical classification of musical instruments on solo recordings“. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, S. V–817–V–820. DOI: 10.1109/ICASSP.2006.1661401.

Eyben, Florian, Sebastian Böck, Björn Schuller und Alex Graves (2010). „Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks.“ In: *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, S. 589–594.

Feldmann, H und W Kumpf (1988). „Listening to music in hearing loss with and without a hearing aid“. In: *Laryngologie, Rhinologie, Otologie* 67.10, S. 489–497.

Fitz, Kelly und Martin McKinney (2015). „Music through hearing aids: perception and modeling“. In: *Proceedings of Meetings on Acoustics*. Bd. 9. 1. Acoustical Society of America. DOI: 10.1121/1.3436580.

Friedman, Jerome, Trevor Hastie und Robert Tibshirani (2001). *The elements of statistical learning*. Bd. 1. Springer series in statistics Springer, Berlin.

Friedrichs, Klaus und Claus Weihs (2012). *Comparing timbre estimation using auditory models with and without hearing loss*. Techn. Ber. 51/2012. Department of Statistics, TU Dortmund University. DOI: 10.17877/DE290R-10355.

Friedrichs, Klaus und Claus Weihs (2013). „Auralization of Auditory Models“. In: *Classification and Data Mining*. Florence, Italy: Springer, S. 225–232. DOI: 10.1007/978-3-642-28894-4_27.

Gfeller, Kate E, Carol Olszewski, Christopher Turner, Bruce Gantz und Jacob Oleson (2006). „Music perception with cochlear implants and residual hearing“. In: *Audiology and Neurotology* 11.Suppl. 1, S. 12–15. DOI: 10.1159/000095608.

Goto, Masataka, Hiroki Hashiguchi, Takuichi Nishimura und Ryuichi Oka (2003). „RWC Music Database: Music genre database and musical instrument sound database.“ In:

Proceedings of the 4th international conference on music information retrieval (ISMIR). Bd. 3, S. 229–230.

Heinz, Michael G, Xuedong Zhang, Ian C Bruce und Laurel H Carney (2001). „Auditory nerve model for predicting performance limits of normal and impaired listeners“. In: *Acoustics Research Letters Online* 2.3, S. 91–96. DOI: 10.1121/1.1387155.

High Performance Computer-Cluster LiDOng. URL: http://lidong.itmc.tu-dortmund.de/ldw/index.php?title=System_overview&oldid=322.

Hines, Andrew und Naomi Harte (2010). „Speech intelligibility from image processing“. In: *Speech Communication* 52.9, S. 736–752. DOI: 10.1016/j.specom.2010.04.006.

Holzapfel, André, Yannis Stylianou, Ali C Gedik und Barış Bozkurt (2010). „Three dimensions of pitched instrument onset detection“. In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.6, S. 1517–1527. DOI: 10.1109/TASL.2009.2036298.

Hornbostel, E. von und C. Sachs (1961). „Classification of musical instruments: Translated from the original german by Anthony Baines and Klaus P. Wachsmann“. In: *The Galpin Society Journal*, S. 3–29. DOI: 10.2307/842168.

Huang, Deng, Theodore T Allen, William I Notz und Ning Zeng (2006). „Global optimization of stochastic black-box systems via sequential kriging meta-models“. In: *Journal of global optimization* 34.3, S. 441–466. DOI: 10.1007/s10898-005-2454-3.

Jensen, Jesper Højvang, Mads Græsbøll Christensen und Søren Holdt Jensen (2007). „A framework for analysis of music similarity measures“. In: *Proc. European Signal Processing Conf*, S. 926–930.

Jepsen, Morten L und Torsten Dau (2011). „Characterizing auditory processing and perception in individual listeners with sensorineural hearing loss“. In: *The Journal of the Acoustical Society of America* 129.1, S. 262–281. DOI: 10.1121/1.3518768.

Jones, Donald R, Matthias Schonlau und William J Welch (1998). „Efficient global optimization of expensive black-box functions“. In: *Journal of Global optimization* 13.4, S. 455–492. DOI: 10.1023/A:1008306431147.

Jürgens, Tim, Stephan D Ewert, Birger Kollmeier und Thomas Brand (2014). „Prediction of consonant recognition in quiet for listeners with normal and impaired hearing using an auditory modela“. In: *The Journal of the Acoustical Society of America* 135.3, S. 1506–1517. DOI: 10.1121/1.4864293.

Karatzoglou, Alexandros, Alex Smola, Kurt Hornik und Achim Zeileis (2004). „kernlab – An S4 Package for Kernel Methods in R“. In: *Journal of Statistical Software* 11.9, S. 1–20. DOI: 10.18637/jss.v011.i09.

Karbasi, M. und D. Kolossa (2015). „A Microscopic Approach to Speech Intelligibility Prediction using Auditory Models“. In: *Proc. Annual Meeting of the German Acoustical Society (DAGA)*, S. 16–19.

Klapuri, Anssi (1999). „Sound onset detection by applying psychoacoustic knowledge“. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Bd. 6. IEEE, S. 3089–3092. DOI: 10.1109/ICASSP.1999.757494.

Klapuri, Anssi (2008). „Multipitch analysis of polyphonic music and speech signals using an auditory model“. In: *Audio, Speech, and Language Processing, IEEE Transactions on* 16.2, S. 255–266. DOI: 10.1109/TASL.2007.908129.

Klapuri, Anssi (2009). „A classification approach to multipitch analysis“. In: *6th Sound and Music Computing Conference, Porto, Portugal*, S. 225–227.

Kohavi, Ron und George H John (1997). „Wrappers for feature subset selection“. In: *Artificial intelligence* 97.1, S. 273–324. DOI: 10.1016/S0004-3702(97)00043-X.

Krige, Daniel G (1951). „A statistical approach to some mine valuation and allied problems on the Witwatersrand“. Diss. University of the Witwatersrand, Faculty of Engineering.

Lang, Michel (2015). „Automatische Modellselektion in der Ueberlebenszeitanalyse“. Diss. Department of Statistics, TU Dortmund University.

Langner, G (1981). „Neuronal mechanisms for pitch analysis in the time domain“. In: *Experimental brain research* 44.4, S. 450–454.

Lartillot, O. und P. Toiviainen (2007). „A MATLAB TOOLBOX FOR MUSICAL FEATURE EXTRACTION FROM AUDIO“. In: *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*, S. 1–8.

Liaw, Andy und Matthew Wiener (2002). „Classification and Regression by randomForest“. In: *R News* 2.3, S. 18–22. URL: http://CRAN.R-project.org/doc/Rnews/.

Ligges, Uwe, Sebastian Krey, Olaf Mersmann und Sarah Schnackenberg (2014). *tuneR: Analysis of music*. URL: http://r-forge.r-project.org/projects/tuner/.

Looi, Valerie, Hugh McDermott, Colette McKay und Louise Hickson (2008). „Music perception of cochlear implant users compared with that of hearing aid users“. In: *Ear and hearing* 29.3, S. 421–434. DOI: 10.1097/AUD.0b013e31816a0d0b.

Lopez-Najera, Alberto, Enrique A Lopez-Poveda und Ray Meddis (2007). „Further studies on the dual-resonance nonlinear filter model of cochlear frequency selectivity: Responses to tones)“. In: *The Journal of the Acoustical Society of America* 122.4, S. 2124–2134. DOI: 10.1121/1.2769627.

Lopez-Poveda, Enrique A und Ray Meddis (2001). „A human nonlinear cochlear filterbank“. In: *The Journal of the Acoustical Society of America* 110.6, S. 3107–3118. DOI: 10.1121/1.1416197.

Madsen, Søren Tjagvad und Gerhard Widmer (2006). „Music complexity measures predicting the listening experience“. In: *In Proc. 9th Int. Conf. Music Perception and Cognition*. URL: http://www.cp.jku.at/research/papers/Madsen_Widmer_ICMPC_2006.pdf.

Martin, Keith D und Youngmoo E Kim (1998). „Musical instrument identification: A pattern-recognition approach“. In: *The Journal of the Acoustical Society of America* 104.3, S. 1768–1768. DOI: 10.1121/1.424083.

Martin, Keith Dana (1999). „Sound-source recognition: A theory and computational model“. Diss. Massachusetts Institute of Technology.

McDermott, Hugh J (2004). „Music perception with cochlear implants: a review“. In: *Trends in amplification* 8.2, S. 49–82. DOI: 10.1177/108471380400800203.

McLeod, Philip (2009). „Fast, accurate pitch detection tools for music analysis“. In: *PhD Thesis, University of Otago. Department of Computer Science*.

Meddis, Ray (2006). „Auditory-nerve first-spike latency and auditory absolute threshold: a computer model“. In: *The Journal of the Acoustical Society of America* 119.1, S. 406–417. DOI: 10.1121/1.2139628.

Meddis, Ray und Michael J Hewitt (1991). „Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification“. In: *The Journal of the Acoustical Society of America* 89.6, S. 2866–2882. DOI: 10.1121/1.400725.

Meddis, Ray und Lowel O'Mard (1997). „A unitary model of pitch perception“. In: *The Journal of the Acoustical Society of America* 102.3, S. 1811–1820. DOI: 10.1121/1.420088.

Meddis, Ray und Lowel P O'Mard (2005). „A computer model of the auditory-nerve response to forward-masking stimuli“. In: *The Journal of the Acoustical Society of America* 117.6, S. 3787–3798. DOI: 10.1121/1.1893426.

Meddis, Ray, Wendy Lecluyse, Christine M Tan, Manasa R Panda und Robert Ferry (2010a). „Beyond the audiogram: Identifying and modeling patterns of hearing deficits“. In: S. 631–640. DOI: 10.1007/978-1-4419-5686-6_57.

Meddis, Ray, Enrique A Lopez-Poveda, Richard R Fay und Arthur N Popper (2010b). *Computational models of the auditory system*. Springer.

Moore, Brian CJ, Brian R Glasberg und Thomas Baer (1997). „A model for the prediction of thresholds, loudness, and partial loudness“. In: *Journal of the Audio Engineering Society* 45.4, S. 224–240.

Mukherjee, Sayan, Pablo Tamayo, Simon Rogers, Ryan Rifkin, Anna Engle, Colin Campbell, Todd R Golub und Jill P Mesirov (2003). „Estimating dataset size requirements for classifying DNA microarray data“. In: *Journal of Computational Biology* 10.2, S. 119–142.

Nagathil, Anil und Rainer Martin (2012). „Optimal signal reconstruction from a constant-Q spectrum“. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, S. 349–352. DOI: 10.1109/ICASSP.2012.6287888.

Nagathil, Anil, Claus Weihs und Rainer Martin (2016). „Spectral Complexity Reduction of Music Signals for Mitigating Effects of Cochlear Hearing Loss“. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.3, S. 445–458. DOI: 10.1109/TASLP.2015.2511623.

Naghatil, A. und R. Martin (2016). „Signal-Level Features“. In: *Music Data Analysis: Foundations and Applications (to appear)*. Hrsg. von C. Weihs, D. Jannach, I. Vatolkin und G Rudolph. Taylor & Francis.

Oxenham, Andrew J (2008). „Pitch perception and auditory stream segregation: implications for hearing loss and cochlear implants“. In: *Trends in amplification* 12.4, S. 316–331.

Panda, Manasa R, Wendy Lecluyse, Christine M Tan, Tim Jürgens und Ray Meddis (2014). „Hearing dummies: Individualized computer models of hearing impairment“. In: *International journal of audiology* 53.10, S. 699–709. DOI: 10.3109/14992027.2014.917206.

Patil, Kailash und Mounya Elhilali (2015). „Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases“. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2015.1, S. 1–13. DOI: 10.1186/s13636-015-0070-9.

Patterson, Roy D, K Robinson, J Holdsworth, D McKeown, C Zhang und M Allerhand (1992). „Complex sounds and auditory images“. In: *Auditory physiology and perception* 83, S. 429–446.

Quinn, B. G. (1994). „Estimating frequency by interpolation using Fourier coefficients“. In: *IEEE Transactions on Signal Processing* 42.5, S. 1264–1268. DOI: 10.1109/78.295186.

Raab, David H (1963). „Backward masking.“ In: *Psychological Bulletin* 60.2, S. 118.

Rihaczek, August W (1968). „Signal energy distribution in time and frequency“. In: *Information Theory, IEEE Transactions on* 14.3, S. 369–374.

Roederer, Juan G und Friedemann Mayer (1999). *Physikalische und psychoakustische Grundlagen der Musik (3. Auflage)*. Springer.

Rosao, Carlos, Ricardo Ribeiro und David Martins De Matos (2012). „Influence of Peak Selection Methods on Onset Detection.“ In: *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, S. 517–522.

Sandrock, T. (2013). „Multi-label feature selection with application to musical instrument recognition“. Diss. Stellenbosch: Stellenbosch University.

Slaney, M. u. a. (1993). „An efficient implementation of the Patterson-Holdsworth auditory filter bank“. In: *Apple Computer, Perception Group, Tech. Rep* 35.

Slaney, Malcolm (1998). „Auditory toolbox“. In: *Interval Research Corporation, Tech. Rep* 10.

Sohn, Wolfgang (2000). „Schwerhörigkeit in Deutschland: Repräsentative Hörscreening-Studie 1999“. In: *DSB-Report/Deutscher Schwerhörigenbund* 3, S. 10–14.

Stein, Michael (1987). „Large sample properties of simulations using Latin hypercube sampling“. In: *Technometrics* 29.2, S. 143–151. DOI: 10.1080/00401706.1987.10488205.

Stevens, S. S., J. Volkmann und E. B. Newman (1937). „A Scale for the Measurement of the Psychological Magnitude Pitch“. In: *The Journal of the Acoustical Society of America* 8.3, S. 185–190. DOI: 10.1121/1.1915893.

Sumner, C. J., L. P. O'Mard, E. A. Lopez-Poveda und R. Meddis (2003). „A nonlinear filter-bank model of the guinea-pig cochlear nerve: rate responses“. In: *The Journal of the Acoustical Society of America* 113.6, S. 3264–3274. DOI: 10.1121/1.1568946.

Terminology, ANSI Psychoacoustical (1973). „S3. 20“. In: *New York, NY: American National Standards Institute*.

Vapnik, V. (1998). *Statistical Learning Theory*. USA: John Wiley und Sons.

Vatolkin, I., M. Preuß, G. Rudolph, M. Eichhoff und C. Weihs (2012). „Multi-objective evolutionary feature selection for instrument recognition in polyphonic audio mixtures“. In: *Soft Computing* 16.12, S. 2027–2047. DOI: 10.1007/s00500-012-0874-9.

Weihs, C., K. Friedrichs und B. Bischl (2012). „Statistics for hearing aids: Auralization“. In: *Second Bilateral German-Polish Symposium on Data Analysis and its Applications (GPSDAA)*, S. 183–196.

Weihs, C., S. Herbrandt, N. Bauer, K. Friedrichs und D. Horn (2016). „Efficient Global Optimization: Motivation, Variation, and Application (accepted)“. In: *Archives of Data Science*. KIT Scientific Publishing.

Wieczorkowska, A., E. Kubera und A. Kubik-Komar (2011). „Analysis of Recognition of a Musical Instrument in Sound Mixes Using Support Vector Machines“. In: *Fundamenta Informaticae* 107.1, S. 85–104.

Wintersohl, K. (2014). *Instrumenten Klassifikation mit Hilfe eines auditorischen Modells*. Bachelor Thesis, Department of Statistics, TU Dortmund University.

Zilany, Muhammad SA und Ian C Bruce (2006). „Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery“. In: *The Journal of the Acoustical Society of America* 120.3, S. 1446–1466. DOI: 10.1121/1.2225512.

A Tabellen

	Mittl. Fehlerrate in %	Max. Fehlerrate in %
1 Partialton, 1 s	0.1	0.3
1 Partialton, 0.2 s	0.1	0.5
1 Partialton, 0.05 s	0.2	0.7
3 Partialtöne, 1 s	0.2	0.7
3 Partialtöne, 0.2 s	0.2	0.6
3 Partialtöne, 0.05 s	0.6	1.3
6 Partialtöne, 1 s, 0prob = 0	0.4	0.8
6 Partialtöne, 0.2 s, 0prob = 0	0.5	1.0
6 Partialtöne, 0.05 s, 0prob = 0	1.3	2.2
6 Partialtöne, 1 s, 0prob = 0.2	0.6	1.2
6 Partialtöne, 0.2 s, 0prob = 0.2	0.6	1.1
6 Partialtöne, 0.05 s, 0prob = 0.2	1.1	1.9
6 Partialtöne, 1 s, 0prob = 0.5	0.5	1.1
6 Partialtöne, 0.2 s, 0prob = 0.5	0.5	1.2
6 Partialtöne, 0.05 s, 0prob = 0.5	1.0	1.8
10 Partialtöne, 1 s, 0prob = 0	0.6	1.4
10 Partialtöne, 0.2 s, 0prob = 0	1.2	2.8
10 Partialtöne, 0.05 s, 0prob = 0	2.6	6.6
10 Partialtöne, 1 s, 0prob = 0.2	1.2	3.0
10 Partialtöne, 0.2 s, 0prob = 0.2	1.4	3.2
10 Partialtöne, 0.05 s, 0prob = 0.2	2.3	5.0
10 Partialtöne, 1 s, 0prob = 0.5	0.9	2.4
10 Partialtöne, 0.2 s, 0prob = 0.5	0.9	2.1
10 Partialtöne, 0.5 s, 0prob = 0.5	1.4	2.7

Tabelle A.1: Mittlere und maximale Fehlerrraten der Partialtonerkennung über alle 30 Kanäle.

Kanalnummer	1	2	3	4	5	6	7	8	9	10
Fehlerrate in %	0.2	0.4	0.5	0.5	0.3	0.3	0.3	0.4	0.6	0.7
Kanalnummer	11	12	13	14	15	16	17	18	19	20
Fehlerrate in %	0.6	0.5	0.7	0.7	0.6	0.9	1.0	1.2	1.4	1.3
Kanalnummer	21	22	23	24	25	26	27	28	29	30
Fehlerrate in %	1.4	1.3	1.4	1.4	1.3	1.4	1.3	1.1	1.0	1.1

Tabelle A.2: Mittlere Fehlerraten der Partialtonerkennung über alle 24 Datensätze.

Klassifikationsaufgabe	SVML	RF
Level 1: Standardmerkmale	0.14	0.14
Level 2: Standardmerkmale	0.18	0.20
Level 1: Ohrmodellmerkmale	**0.08**	0.10
Level 2: Ohrmodellmerkmale	**0.14**	0.15

Tabelle A.3: Fehlerraten der 2 Knoten der hierarchischen Taxonomie mit Instrumentenfamilien.

Klassifikationsaufgabe	SVML	RF
Level 1: Standardmerkmale	0.08	0.06
Level 2: Standardmerkmale	0.23	0.18
Level 3: Standardmerkmale	**0.08**	0.14
Level 1: Ohrmodellmerkmale	**0.04**	0.05
Level 2: Ohrmodellmerkmale	**0.12**	0.14
Level 3: Ohrmodellmerkmale	0.11	0.13

Tabelle A.4: Fehlerraten der 3 Knoten der Hornbostel-Sachs Taxonomie.

	a		**b**		**c**	
Anpassung	$R^2 = 0.26, R_a^2 = 0.07$		$R^2 = 0.60, R_a^2 = 0.50$		$R^2 = 0.66, R_a^2 = 0.57$	
Faktoren	Schätzer	p-Wert	Schätzer	p-Wert	Schätzer	p-Wert
(Intercept)	0.8297	**<2e-16**	0.6466	**<2e-16**	0.6705	**<2e-16**
Intervall	0.0077	0.55	0.0089	0.57	0.0357	0.11
Begl.-Einsätze	-0.0037	0.78	-0.0634	**3e-04**	-0.0421	**0.06**
Dynamik	-0.0172	0.19	-0.0047	0.76	0.0054	0.81
Begl.-Instrument	0.0253	**0.06**	-0.0194	0.22	-0.1459	**3e-07**
Tonhöhe	-0.0117	0.37	0.0479	**4e-03**	0.0235	0.29
Tondauer	-0.0131	0.31	-0.0424	**0.01**	-0.0025	0.91
Begl.-Tonhöhe	0.0156	0.22	-0.0394	**0.02**	-0.0288	0.20

Tabelle A.5: Auswertung über alle Instrumente und alle Plackett-Burman-Designs für ***Hearing Dummy* 2**. Die Zielvariable ist der mittlere F-Wert – **a**: monophone Einsatzzeiterkennung, **b**: predominante Einsatzzeiterkennung und **c**: polyphone Einsatzzeiterkennung (**dick** = signifikant zum 10%-Level).

Anpassung	**a** $R^2 = 0.27, R_a^2 = 0.08$		**b** $R^2 = 0.58, R_a^2 = 0.48$		**c** $R^2 = 0.66, R_a^2 = 0.57$	
Faktoren	Schätzer	p-Wert	Schätzer	p-Wert	Schätzer	p-Wert
(Intercept)	0.8251	**<2e-16**	0.6668	**<2e-16**	0.6826	**<2e-16**
Intervall	0.0016	0.90	-0.0000	1.00	0.0283	0.19
Begl.-Einsätze	-0.0067	0.62	-0.0524	**7e-04**	-0.0474	**0.03**
Dynamik	-0.0263	**0.06**	-0.0133	0.42	-0.0053	0.80
Begl.-Instrument	0.0240	**0.08**	-0.0185	0.19	-0.1314	**9e-07**
Tonhöhe	-0.0084	0.53	0.0551	**4e-04**	0.0200	0.35
Tondauer	-0.0151	0.26	-0.0268	**0.06**	-0.0020	0.93
Begl.-Tonhöhe	0.0127	0.34	-0.0198	0.16	-0.0251	0.24

Tabelle A.6: Auswertung über alle Instrumente und alle Plackett-Burman-Designs für ***Hearing Dummy* 3**. Die Zielvariable ist der mittlere F-Wert – **a**: monophone Einsatzzeiterkennung, **b**: predominante Einsatzzeiterkennung und **c**: polyphone Einsatzzeiterkennung (**dick** = signifikant zum 10%-Level).

B Grafiken

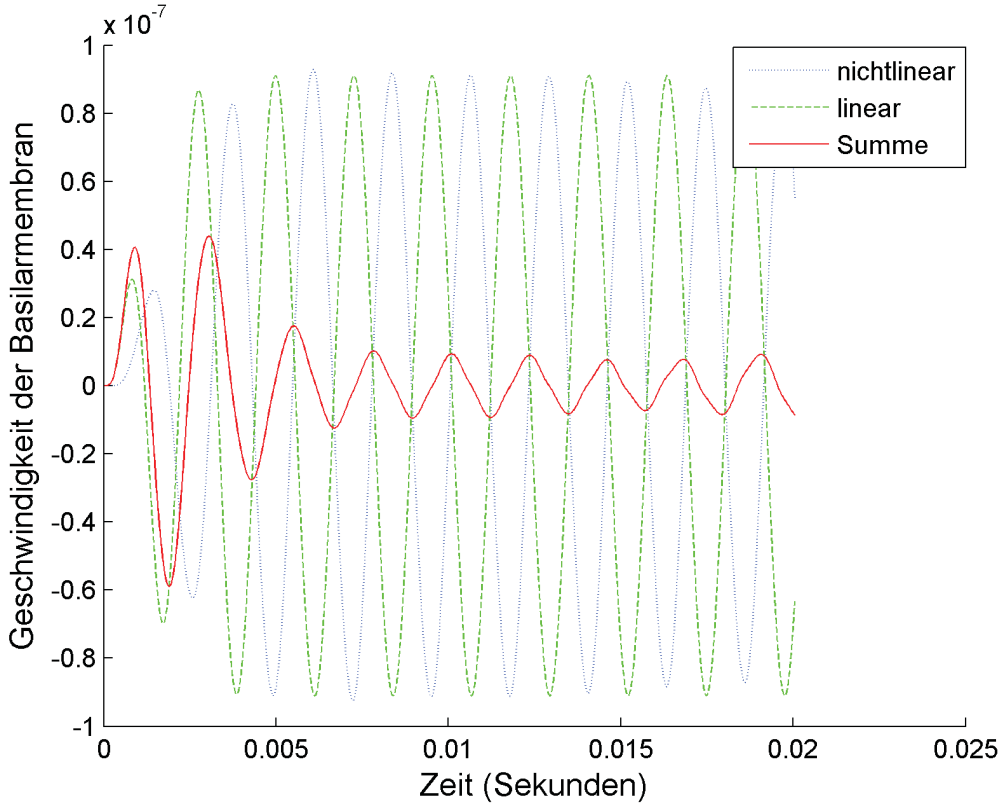

Abbildung B.1: Geschwindigkeit der Basilarmembran für Kanal 13 bei einem reinen Ton von 440 Hz und einer Lautstärke von 60 dB SPL. Bei dieser speziellen Konstellation heben sich durch eine Phasenverschiebung von etwa 180° der lineare und der nichtlineare Pfad fast vollständig gegenseitig auf.

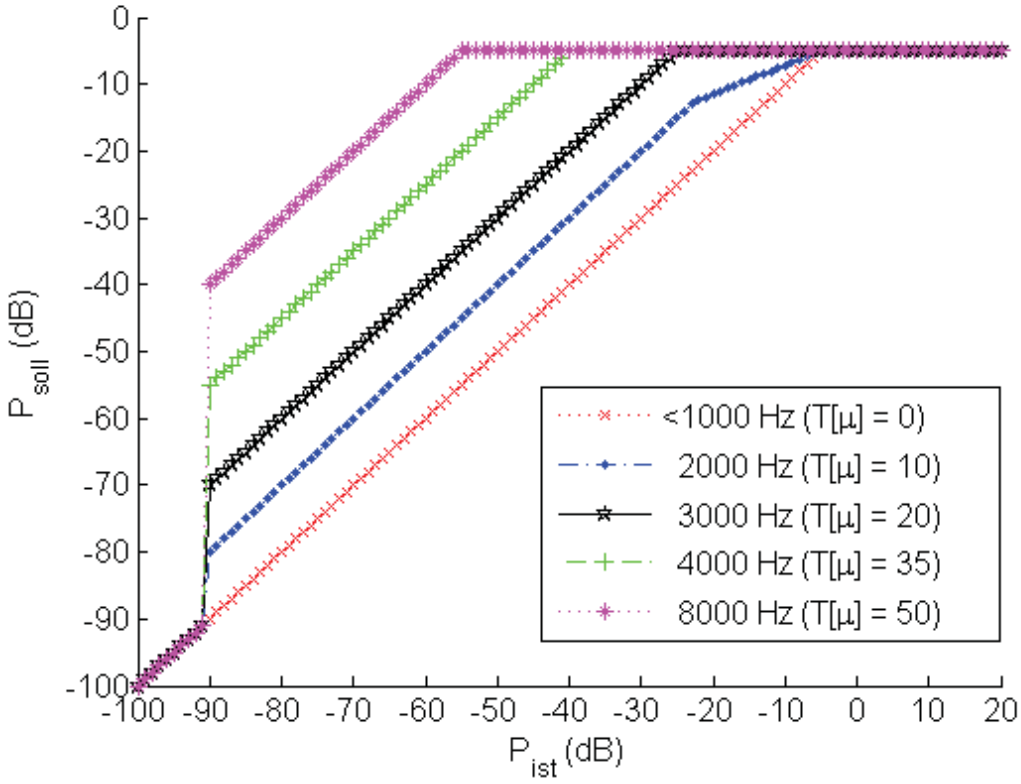

Abbildung B.2: Ergebnis der Optimierung des Kniepunktes: Zusammenhang zwischen vorliegendem Pegel P_{ist} und gewünschtem Pegel P_{soll} in Abhängigkeit von der Frequenz (bzw. dem durch die Experteneinstellung entsprechenden Verstärkungsfaktor $T[\mu]$).