Notentext-Informierte Quellentrennung für Musiksignale

Meinard Müller[1], Jonathan Driedger[1], Sebastian Ewert[2]

[1]International Audio Laboratories Erlangen*

meinard.mueller@audiolabs-erlangen.de, jonathan.driedger@audiolabs-erlangen.de

[2]Queen Mary, University of London

sebastian.ewert@eecs.qmul.ac.uk

Abstract: Die automatisierte Zerlegung von Musiksignalen in elementare Bestandteile stellt eine zentrale Aufgabe im Bereich der Musikverarbeitung dar. Hierbei geht es unter anderem um die Identifikation und Rekonstruktion von individuellen Melodie- und Instrumentalstimmen aus einer als Wellenform gegebenen Audioaufnahme – eine Aufgabenstellung, die im übergeordneten Bereich der Audiosignalverarbeitung auch als Quellentrennung bezeichnet wird. Im Fall von Musik weisen die Einzelstimmen typischer Weise starke zeitliche und spektrale Überlappungen auf, was die Zerlegung in die Quellen ohne Zusatzwissen zu einem im Allgemeinen kaum lösbaren Problem macht. Zur Vereinfachung des Problems wurden in den letzten Jahren zahlreiche Verfahren entwickelt, bei denen neben dem Musiksignal auch die Kenntnis des zugrundeliegenden Notentextes vorausgesetzt wird. Die durch den Notentext gegebene Zusatzinformation zum Beispiel hinsichtlich der Instrumentierung und den vorkommenden Noten kann zur Steuerung des Quellentrennungsprozesses ausgenutzt werden, wodurch sich auch überlappende Quellen zumindest zu einem gewissen Grad trennen lassen. Weiterhin lassen sich durch den Notentext die zu trennenden Stimmen oft erst spezifizieren. In diesem Artikel geben wir einen Überblick über neuere Entwicklungen im Bereich der Notentext-informierten Quellentrennung, diskutieren dabei allgemeine Herausforderungen bei der Verarbeitung von Musiksignalen, und skizzieren mögliche Anwendungen.

1 Einleitung

Die Zerlegung von überlagerten Schallquellen in ihre Einzelbestandteile, auch als Quellentrennung ("Source Separation") bekannt, stellt eine der zentralen Fragestellungen der digitalen Audiosignalverarbeitung dar. Im Bereich der digitalen Sprachsignalverarbeitung geht es zum Beispiel in dem als "Cocktail Party Scenario" bekannten Problem darum, aus einer Audioaufnahme von mehreren gleichzeitig redenden Sprechern die einzelnen Sprachsignale zu rekonstruieren [Che53]. Auch im Bereich der Musiksignalverarbeitung gibt es zahlreiche verwandte Fragestellungen, die häufig unter dem Begriff der Quellentrennung subsumiert werden. Hierbei entsprechen den Quellen gewisse Melodie- oder Instrumentalstimmen, die es aus einem polyphonen Klanggemisch herauszutrennen

*Die International Audio Laboratories Erlangen sind eine gemeinsamen Einrichtung der Friedrich-Alexander-Universität Erlangen-Nürnberg und des Fraunhofer-Instituts für Integrierte Schaltungen IIS.

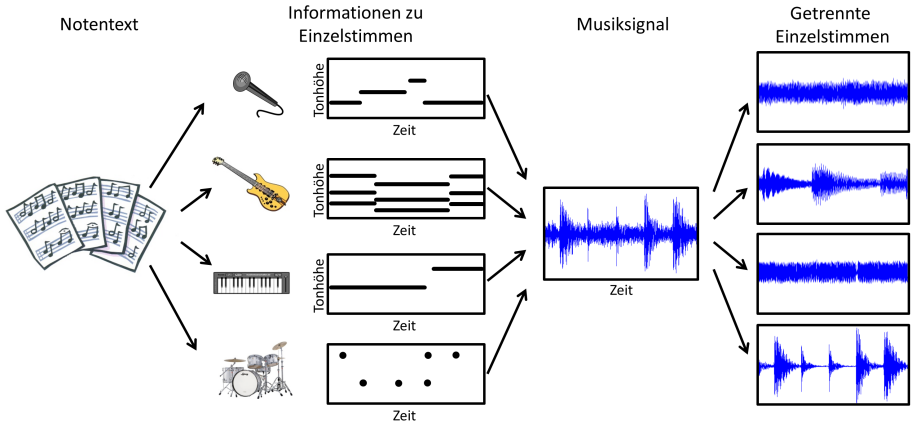

Abbildung 1: Notentext-informierte Quellentrennung.

gilt [EM12a, GR08, Got00, MEKR11, PEE+07, RK08]. Viele Verfahren zur Quellentrennung basieren auf Annahmen wie dem Vorliegen mehrerer Audiokanäle (zum Beispiel mehrere Mikrophonsignale aufgenommen aus verschiedenen Richtungen) oder der statistischen Unabhängigkeit der Quellensignale. Im Fall von Musik sind diese Annahmen allerdings häufig nicht zutreffend. Zum einen liegen Musikaufnahmen oft nur in Mono (ein Kanal) oder in Stereo (zwei Kanäle) vor. Zum anderen sind die verschiedenen musikalischen Quellen einer Aufnahme in den meisten Fällen nicht unabhängig. Ganz im Gegenteil sind Instrumentalstimmen in polyphonen Audioaufnahmen typischerweise stark korreliert: Sie teilen dieselben Harmonien, folgen denselben melodischen Linien und Rhythmen und interagieren miteinander. Dies macht die Rekonstruktion von musikalischen Stimmen aus einem polyphonen Klanggemisch zu einem äußerst schwierigen und im Allgemeinen auch unlösbaren Problem.

Bei der Zerlegung von Musiksignalen werden daher häufig musikalische Eigenschaften und weiteres Zusatzwissen genutzt. Zum Beispiel zeichnet sich die Melodiestimme häufig durch Dominanz in der Lautstärke und zeitliche Kontinuität aus, was ihre Extraktion wesentlich erleichtern kann [Bre90, Dre11, SG12]. Weiterhin wird bei der Trennung der Basslinie ausgenutzt, dass es sich hierbei meist um die tiefste Stimme handelt [Got04]. Die Extraktion der Schlagzeugspur lässt sich inbesondere dann gut bewerkstelligen, wenn die anderen Quellen vorwiegend harmonischer Natur sind. Hierbei wird ausgenutzt, dass perkussive Elemente (vertikale spektrale Strukturen) sich grundsätzlich von harmonischen Elementen (horizontale spektrale Strukturen) unterscheiden [OMKS08, Fit10]. Weiterhin lässt sich eine Singstimme häufig von den anderen Begleitstimmen dadurch abgrenzen, dass sie ein starkes Vibrato und Gleiteffekte aufweist [RP09].

In den letzten Jahren wurden auch verstärkt multimodale Strategien der Quellentrennung entwickelt, bei denen unter anderem die Kenntnis des Notentexts ausgenutzt wird (“Score-Informed Source Separation”), siehe Abbildung 1. Neuere Ansätze zeigen, dass sich mittels dieser Zusatzinformation auch stark überlappende Quellen zu einem gewissen Grad

trennen lassen [EM12b, HDB11, IGK+08, WPD06]. Weiterhin lassen sich über die Notenschrift die zu trennenden Stimmen leicht spezifizieren und dann auf die Audiodomäne übertragen. Die Strategie einer informierten Quellentrennung setzt allerdings voraus, dass der Notentext synchronisiert zu den Audiodaten vorliegt. Die automatisierte Berechnung einer zeitlich hochauflösenden Synchronisation stellt allerdings für sich schon ein schwieriges Forschungsproblem dar [EMG09, JER11].

Mit diesem Artikel verfolgen wir im Wesentlichen zwei Ziele. Zum einen sollen anhand des Quellentrennungsproblems allgemeine Herausforderungen diskutiert werden, denen man sich bei der Verarbeitung von Musiksignalen stellen muss. Zum anderen soll ein Überblick über neuere Entwicklungen im Bereich der Notentext-informierten Quellentrennung gegeben werden, wobei weniger die technischen Details als vielmehr eine anschauliche Darstellung der zugrundeliegenden Ideen im Fokus steht. Nach dieser Einleitung setzen wir unsere Diskussion mit der Frage fort, warum Musiksignale komplex sind (Abschnitt 2). Selbst für ein- und dasselbe Musikstück können ganz unterschiedliche Darstellungformen existieren wie zum Beispiel der Notentext und unterschiedliche Audioaufnahme. Wir gehen daher auf die Aufgabenstellung der Musiksynchronisation ein, bei der es um die Verlinkung unterschiedlicher Darstellungsformen von Musik geht (Abschnitt 3). Durch synchronisiert vorliegende Notentextinformationen können dann Musikanalyseaufgaben unterstützt oder gar erst ermöglicht werden. Wir diskutieren diese Strategie anhand von drei Szenarien im Bereich der Quellentrennung. Zum einen betrachten wir parametrische Modelle (Abschnitt 4) und Matrixfaktorisierungsverfahren (Abschnitt 5) zur notenbasierten Parametrisierung von Musiksignalen. Zum anderen zeigen wir, wie Notentextinformationen zur Schätzung der Fundamentalfrequenz und Abtrennung einer Singstimme verwendet werden können (Abschnitt 6). Der Artikel schließt mit der Diskussion einiger Anwendungen und einem kurzen Fazit (Abschnitt 7).

2 Warum sind Musiksignale komplex?

Musik ist ein allgegenwärtiger Teil unseres Lebens und kann uns durch ihre emotionale Kraft beruhigen, aufwühlen und auf überraschende und tiefgreifende Weise berühren [MEKR11]. Musik bietet eine enorme Bandbreite an Formen und Stilen, angefangen von einfachen, unbegleiteten Volkslieder, über orchestrale Werke, bis hin zu minutiös konstruierten Stücken elektronischer Musik. Bei der Verarbeitung von Musikdaten spielen ganz unterschiedliche musikalische Aspekte wie die Rhythmik, Dynamik, Harmonik oder Klangfarbe eine Rolle. Diese Aspekte wiederum können auf unterschiedlichen zeitlichen Stufen betrachtet werden und führen zu komplexen Hierarchien musikalischer Strukturen. Weiterhin kann man in der Musik unterschiedliche Grade der Mehrstimmigkeit betrachten, angefangen von einstimmiger Musik (Monophonie), über mehrstimmige Musik mit einer melodischen Hauptstimme, die von anderen Instrumenten begleitet wird (Homophonie), bis hin zu komplexer mehrstimmiger Musik mit mehreren unabhängigen Stimmen (Polyphonie).

Als Beispiel zeigt Abbildung 2a den Notentext eines polyphonen Klavierstücks. Das im 3/4-Takt gehaltene Stück besitzt eine Hauptmelodiestimme (rechte Hand) und eine Be-

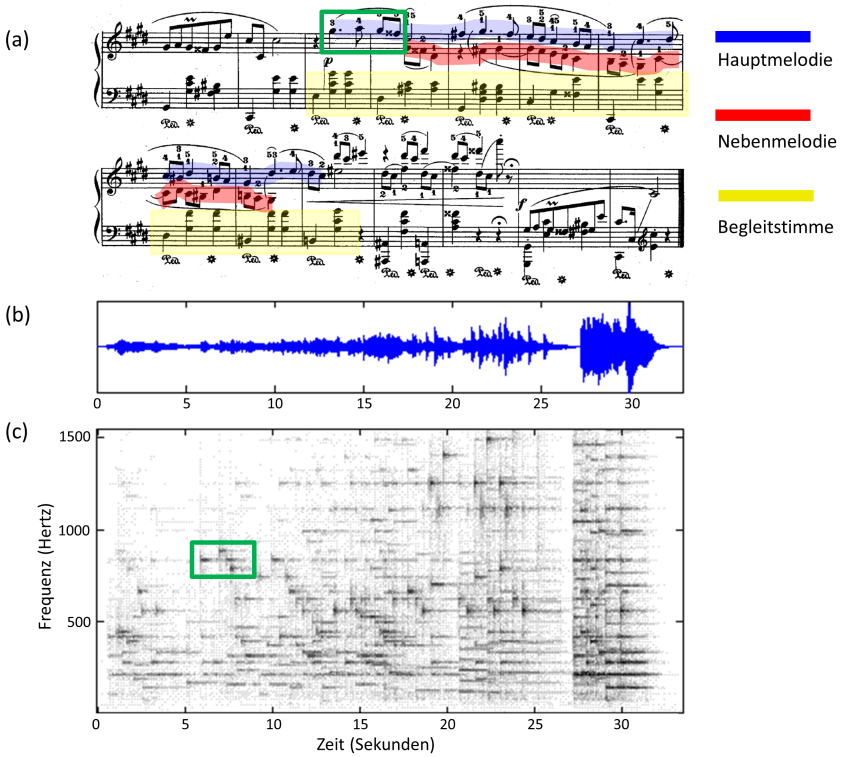

Abbildung 2: **(a)** Verschiedene musikalische Stimmen in einem polyphonen Klavierstück (Ende von Chopins Mazurka Op. 63, Nr. 2). **(b)** Wellenform einer zugehörgen Audioaufnahme (Musiksignal). **(c)** Spektrogramm.

gleitstimme (linke Hand). Zudem setzt in den letzten Takten des Stücks eine weitere Nebenstimme eine, die eine um eine Oktave nach unten versetzte und einen Taktschlag nach hinten verschobene Kopie der Hauptstimme ist, wodurch ein faszinierendes Stimmengeflecht entsteht.

Dieses Beispiel soll die musikalische Komplexität andeuten wie sie schon bei einem einzigen Instrument entstehen kann. Noch vielschichtiger kann die Situation bei orchestraler Musik werden. Darüber hinaus kann die Notentextdarstellung der Musik auf ganz unterschiedliche Weisen interpretiert werden. Notenparameter wie zum Beispiel Tonhöhen, Tondauern oder Einsatzzeiten sind zwar explizit im Notentext gegeben, lassen dem Musiker aber zumeist Spielraum hinsichtlich des Tempos, der Dynamik oder der Ausführung von Notengruppen.

Geht man vom Notentext zu den Audiosignalen über, so wird die Analyse der Daten meist noch komplizierter. Bei Musiksignalen hat man es mit *akustischen Wellenformen* zu tun, bei denen Noteninformationen nicht unmittelbar ablesbar sind, siehe Abbildung 2b. Die eindimensionale Wellenform kodiert die relativen Luftdruckschwankungen wie sie vom

Instrument erzeugt werden und schließlich unser Ohr erreichen. Nicht zuletzt können akustische Eigenschaften wie Hall oder Raumklang und die in der Studioproduktion vorgenommenen Modifikationen erheblichen Einfluss auf die Audioaufnahmen haben.

Zur Verarbeitung von Musiksignalen werden die Wellenformen in einem ersten Schritt in geeignete Merkmalsdarstellungen überführt. Prominentestes Beispiel einer solchen Merkmalsdarstellung ist das Spektrogramm, welches über eine gefensterte Fouriertransformation berechnet wird, siehe Abbildung 2c. Eine solche Zeit-Frequenz Darstellung gibt die lokale Energieverteilung des Signals aufgeschlüsselt nach Frequenzbändern an. Auch wenn man in einer solcher Darstellung die Notenereignisse oft schon durch auffallende vertikale und horizontale Strukturen erahnen kann (siehe zum Beispiel die grün umrahmten Bereiche in Abbildung 2), ist die Rekonstruktion der Noten aus einem Spektrogramm für komplexe Musiksignale ein äußerst schwieriges Problem.[1] Ein Grund hierfür ist, dass beim Spielen selbst einer einzelnen Note auf einem Instrument schon ein komplexes Klanggemisch entstehen kann. Beim Klavier kann dieses Klanggemisch zum Beispiel durch den Tastenanschlag hervorgerufene perkussive Strukturen wie auch durch Obertöne hervorgerufene harmonische Strukturen beinhalten. Eine einzelne Note hat damit Auswirkungen auf ganz unterschiedliche Bereiche im Spektrogramm. Bei polyphonen Musiksignalen entstehen daher komplexe zeitlich und spektral sich überlappende Muster, die aufzulösen eine Kernaufgabe der Quellentrennung darstellt.

3 Synchronisation

Aufgrund der oben dargestellten Komplexität von Musiksignalen ist eine Zerlegung des Signals oder des zugehörigen Spektrogramms in Elementarbestandteile ohne Zusatzinformation nur schwer möglich. Ist jedoch zusätzlich ein Notentext vorhanden, so kann man versuchen, den dort spezifizierten Noten geeignete zeitlich-spektrale Muster im Spektrogramm zuzuordnen. Ein erster Schritt hierfür ist die Zuordnung der Notenereignisse zu musikalisch entsprechenden Zeitpunkten in der Audioaufnahme. Dies ist genau das Ziel der *Musiksynchronisation*, bei der es allgemein gesprochen um die automatische Verlinkung zweier Musikdatenströme unterschiedlicher Formate geht, siehe auch Abbildung 3.

Die meisten Verfahren zur Musiksynchronisation gehen in zwei Schritten vor [Mül07]. Im ersten Schritt werden die zu verlinkenden Datenströme in geeignete Merkmalsdarstellungen umgewandelt, um hierdurch zum einen eine Datenreduktion und zum anderen Robustheit gegenüber nicht zu berücksichtigenden Variabilitäten zu erzielen. Im Musikkontext werden insbesondere *Chromamerkmale* mit großem Erfolg für unterschiedliche Retrieval- und Analyseaufgaben eingesetzt [BW05, Mül07]. Diese Merkmale korrelieren stark mit dem Harmonieverlauf des zugrundeliegenden Musikstücks und weisen einen hohen Grad an Robustheit gegenüber Änderungen in Instrumentierung, Klangfarbe und Dynamik auf. Insbesondere eignen sich chromabasierte Merkmale als gemeinsame Mid-Level Darstellung für sowohl akustische als auch symbolische Musikrepräsentationsformen und erlauben damit eine Verlinkung multimodal vorliegender Versionen. Im zweiten Schritt werden

[1] Dieses Problem wird oft auch als Musiktranskription bezeichnet.

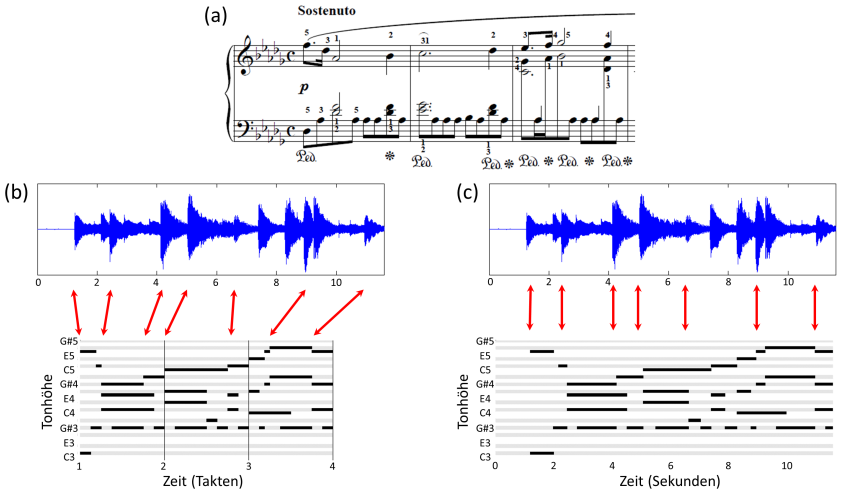

Abbildung 3: Synchronisation von Notentext und Musiksignal (Anfang von Chopins Op. 28, Nr. 15). **(a)** Notentext. **(b)** Wellenform und Klavierwalzendarstellung mit Verlinkung (rote Pfeile). **(c)** Wellenform mit synchroner Klavierwalzendarstellung.

dann die beiden extrahierten Merkmalsfolgen mittels Alignment-Verfahren wie dem *Dynamic Time Warping* (DTW) synchronisiert.

In Abbildung 3b ist ein solches Synchronisationsergebnis mittels der roten Pfeile dargestellt. Der Notentext wurde dabei auf eine sogenannten *Klavierwalzendarstellung* reduziert, bei der die Noten durch geeignete Rechtecke in einem Zeit-Tonhöhen-Raster repräsentiert werden. Das Synchronisationsergebnis erlaubt es nun, die Klavierwalzendarstellung so zeitlich zu verzerren, dass diese synchron zur Audioaufnahme verläuft. Wie wir in den nächsten Abschnitten sehen werden kann die so synchronisierte Klavierwalzendarstellung in gewisser Weise als eine erste Approximation einer Zeit-Frequenz Darstellung der Audioaufnahme angesehen werden.

4 Parametrische Modelle

Viele der in der Literatur beschriebenen Verfahren zur Notentext-informierten Quellentrennung basieren auf sogenannten *parametrischen Modellen*, bei denen die akustischen und musikalischen Eigenschaften des Musiksignals durch geeignete Parameter *explizit* abgebildet werden [EM11, HR07, HDB11, IGK+08, WVR+11]. Der Klang einer einzelnen Note wird zum Beispiel durch Parameter erfasst, die die Tonhöhe und deren zeitlichen Verlauf (z. B. bei Vibrato), die spektrale Hüllkurve und die Obertonzusammensetzung (welche zur Klangfarbe korrelieren) oder den Dynamikverlauf (also die Lautstärke) beschreiben. Andere Parameter können übergeordnete Aspekte wie das Tempo oder den harmonischen Kontext reflektieren.

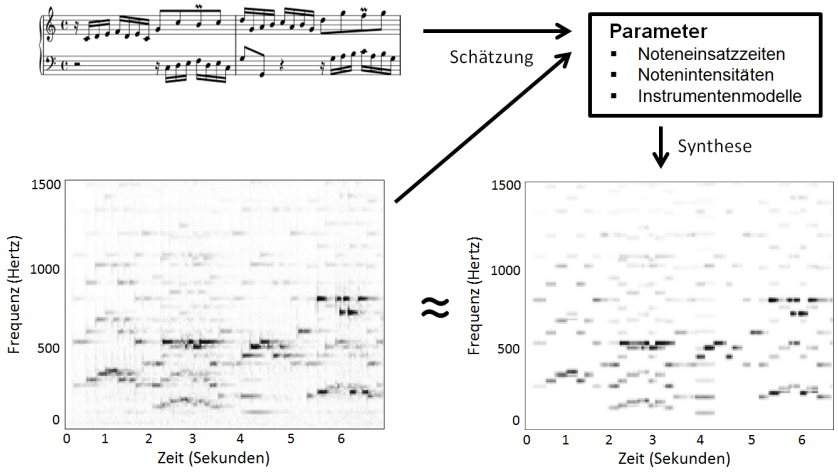

Abbildung 4: Spektrogramm eines Musiksignals (links) und eine über ein parametrisches Modell synthetisierte Approximation (rechts).

Ausgehend von einer Notentextdarstellung und einer Audioaufnahme besteht das Vorgehen vieler Verfahren darin, das Spektrogramm des Musiksignals durch eine kleine Anzahl von explizit gegebenen Parameter zu beschreiben. Dies gelingt natürlich im Allgemeinen nur in einem approximativen Sinne. Bei der Schätzung der Parameter wird meist iterativ vorgegangen, wobei die Synchronisation der aus dem Notentext generierten Klavierwalzendarstellung oft den ersten Schritt darstellt. Insbesondere können hierdurch die zu Einsatzzeiten und Tonhöhen korrespondierenden Parameter geeignet initialisiert werden. In den nächsten Schritten werden dann weitere, die Lautstärken und Klangfarben betreffende Parameter geschätzt. Hierbei werden aus den Parametern Spektraldarstellungen synthetisiert und mit dem Originalspektrogramm verglichen, siehe auch Abbildung 4. Der Abstand wird dann mittels geeigneten Optimierungsverfahren iterativ verkleinert bis eine Konvergenz erreicht wird.

Parametrische Modellen haben den Vorteil, dass viele der zeitlich-spektralen Muster im Spektrogramm durch musikalisch und akustisch interpretierbare Parameter erfasst werden. Mittels dieser Parametrisierung können dann die zu den unterschiedlichen Quellen gehörigen Muster aus dem Spektrogramm herausgetrennt und durch eine inverse Fouriertransformation in Wellenformdarstellungen transformiert werden. Da parametrische Modelle den Suchraum im Allgemeinen relativ stark einschränken, sind die darauf basierenden Schätzungen vergleichsweise robust. Auf der anderen Seite können unzureichende Modellannahmen zu schlechten Approximationsergebnissen und nutzlosen Parametrisierungen führen.

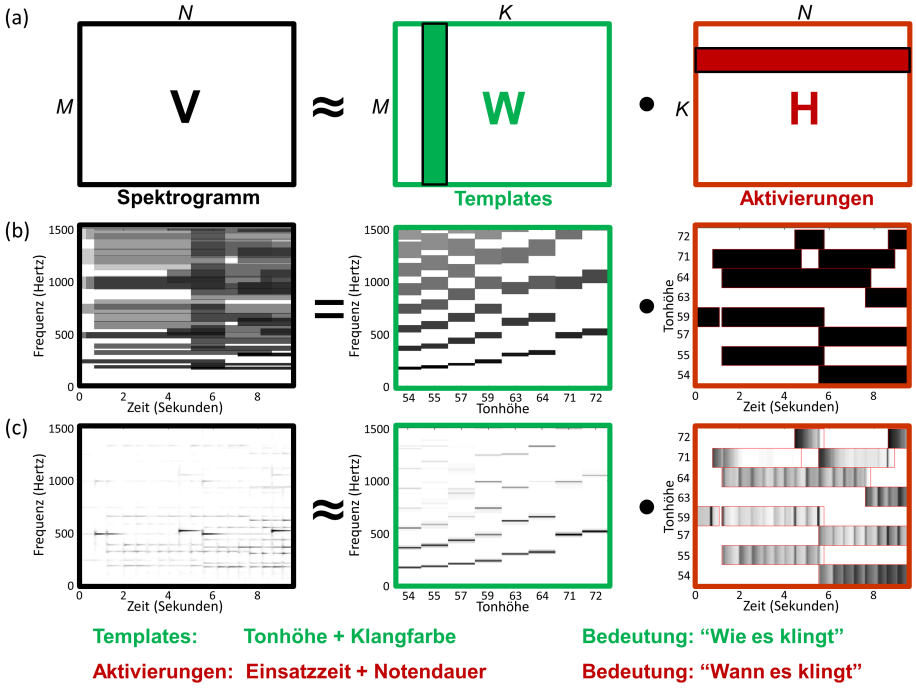

Abbildung 5: NMF-basierte Matrixzerlegung eines Magnituden-Spektrogramms einer Audioaufnahme. **(a)** Schematische Darstellung. **(b)** Zerlegung nach Notentext-basierter Initialisierung. **(c)** Zerlegung nach NMF-Verfeinerung.

5 NMF-basierte Spektrogrammfaktorisierung

Neben parametrischen Modellen wurden in den letzten Jahren verstärkt auch Techniken der Matrixfaktorisierung für die Quellentrennung eingesetzt. Hierbei wurde insbesondere auf eine als NMF (“Non-Negative Matrix Factorization”) bekannte Variante mit der zusätzlichen Forderung, dass bei allen beteiligten Matrizen die Einträge nicht-negativ sind, zurückgegriffen [LS00, SRS08]. Angewendet wird diese Technik auf ein Magnituden-Spektrogramm $V \in \mathbb{R}_{\geq 0}^{M \times N}$, bei dem die komplexen Koeffizienten des Spektrogramms durch ihre Absolutwerte ersetzt werden. Ziel der NMF ist es, diese Matrix in zwei nicht-negative Matrizen $W \in \mathbb{R}_{\geq 0}^{M \times K}$ und $H \in \mathbb{R}_{\geq 0}^{K \times N}$ zu zerlegen, so dass $V \approx W \cdot H$ gilt, siehe Abbildung 5a. Im Musikkontext werden die Spalten von W häufig auch als *Templates* bezeichnet und die Zeilen von H als die zugehörigen Aktivierungen. Intuitiv repräsentieren die Templates die Tonhöhen und Klangfarben der unterschiedlichen, in dem Musikstück vorkommenden Tönen, während die Aktivierungen die Einsatzzeiten und Dauern dieser Töne wiedergeben. Mit anderen Worten kodieren die Templates *wie* etwas klingt, während die Aktivierungen beschreiben *wann* etwas klingt. Hierbei erinnert die Aktivierungsmatrix sehr an die Klavierwalzendarstellung eines Notentextes.

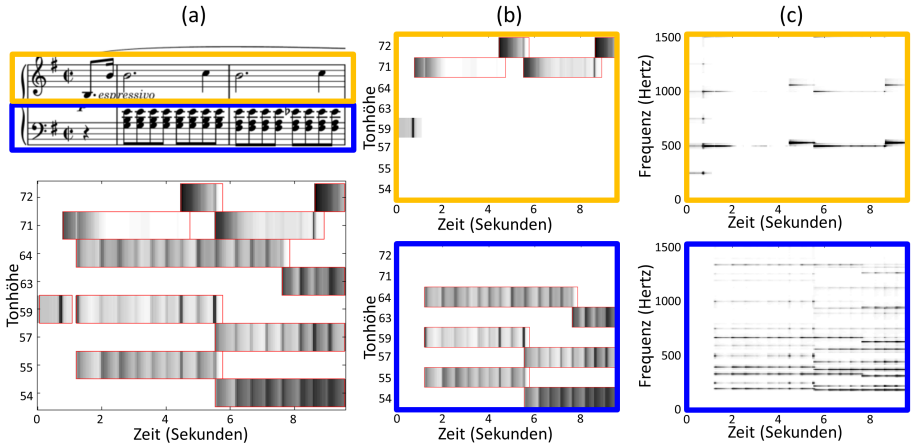

Abbildung 6: Trennung einer Klavieraufname in linke und rechte Hand mittels Notentext-informierter Initialisierung und NMF-basierter Verfeinerung. **(a)** Notentext und gelernte Aktivierungsmatrix (notenweise Initialisierungsbereiche entsprechen roten Umrandungen). **(b)** Aufspaltung in Aktivierungsmatrizen für linke (unten) und rechte (oben) Hand. **(c)** Magnituden-Spektrogramme erhalten durch Multiplikation der jeweiligen Aktivierungsmatrix mit der Templatematrix.

Im Allgemeinen werden NMF-basierte Matrixfaktorisierungen durch iterative Verfahren berechnet, die zu einem lokalen Optimum führen [LS00]. Ohne weitere Einschränkungen entbehrt das Fakorisierungsergebnis oft jeglicher Semantik. Um musikalisch sinnvolle Zerlegungen zu erhalten, wird daher zusätzliches Wissen eingebracht. Zum Beispiel werden den Templates mittels parametrischer Modelle musikalische Eigenschaften bezüglich Tonhöhen und Obertoneigenschaften aufgezwungen [HKV09, HBD10, WVR$^+$11]. Als Alternative können den Templates harmonische Strukturen durch eine geeignete Initialisierung auferlegt werden, wobei alle nicht-relevanten Einträge auf Null und alle relevanten Einträge auf Eins gesetzt werden [ROS07]. Da bei der iterativen Berechnung der NMF-Zerlegen nur multiplikative Update-Regeln zum Einsatz kommen, behalten alle mit Null initialisierten Einträge ihren Wert.

In [EM12b] wurde diese Idee erweitert, indem nicht nur die Templates sondern auch die Aktivierungen über geeignete Binärwerte initialisiert wurden. Unter Hinzunahme eines Notentexts werden hierzu in einem ersten Schritt zu der Audioaufnahme synchrone Klavierwalzendarstellungen berechnet (siehe Abschnitt 3). Die Aktivierungsmatrix wird anhand dieser Klavierwalzendarstellung unter Zulassung gewisser Toleranzbereiche initialisiert, siehe Abbildung 5b für eine Illustration. Weiterhin wird für jede in dem Musikstück vorkommende Tonhöhe ein Template mit einem sehr großzügigem Obertonmodell initialisiert. In gewisser Weise entspricht diese Initialisierung schon grob der gewünschten Matrixfaktorisierung. Durch den anschließenden NMF-Schritt wird diese Faktorisierung nun verfeinert, siehe Abbildung 5c. Insbesondere werden durch die gelernten Aktivierungen die Noteneinsatzzeiten, Intensitäten und Notendauern wiedergegeben, während durch die gelernten Templates die Stimmung und der Klang der Noten erfasst werden.

Ein entscheidender Aspekt bei diesem Verfahren ist, dass bei der NMF-basierten Verfeinerung die Zuordnung der durch den Notentext gegebenen Noten und den ursprünglichen Initialisierungsbereichen[2] bestehen bleibt. Hierdurch erhält man wie schon bei den parametrischen Modellen (Abschnitt 4) eine notenweise Parametrisierung des Magnituden-Spektrogramms. Durch eine solche Parametrisierung kann zum Beispiel eine Klavieraufnahme in zwei Quellen zerlegt werden, die jeweils den von der linken und der rechten Hand gespielten Noten entsprechen, siehe Abbildung 6. Hierbei wird aus der parametrisierten Matrixzerlegung ein Magnituden-Spektrogramm für die linke und die rechte Hand zusammengesetzt. Unter Verwendung der Phaseninformation des Originalspektrogramms werden dann die Wellenformen der jeweiligen Quellen durch Anwendung einer inversen gefensterten Fouriertransformation gewonnen.

Ein wesentlicher Vorteil von NMF-basierten Verfahren ist ihre leichte Implementierbarkeit und hohe Recheneffizienz. Im Gegensatz zu parametrischen Modellen, die abhängig von der mathematischen Modellierung oft komplizierte und rechenintensive Update-Regeln erfordern, werden bei der NMF-Berechnung (auch bei Verwendung von über die Nulleinträge definierten Einschränkungen) nur einfache, multiplikative Update-Regeln benötigt, siehe [EM12a] für weitere Details und Erweiterungen NMF-basierter Verfahren.

6 F0-basierte Abtrennung der Gesangsstimme

Das vorgestellte NMF-basierte Verfahren zur Quellentrennung liefert gute Ergebnisse, wenn sich das zugrundeliegende Spektrogramm durch eine kleine Anzahl spektraler Templates erklären lässt. Dies ist zum Beispiel bei Klaviermusik der Fall, da sich hier jeder Ton schon relativ gut durch ein einziges spektrales Template (bis auf Skalierung) beschreiben lässt. In der Regel ist dies allerdings bei anderen Instrumenten und bei der menschlichen Stimme nicht der Fall. Hier können im zeitlichen Verlauf erhebliche Frequenzschwankungen wie zum Beispiel Frequenzmodulationen beim Vibrato oder kontinuierliche Veränderung der Tonhöhe beim Glissando auftreten. In einigen Verfahren werden, wie in Abschnitt 4 erwähnt, parametrische Modelle verwendet, um Frequenzmodulationen in den Griff zu bekommen. Eine andere Vorgehensweise besteht darin, zunächst den exakten Verlauf der Grundfrequenz (F0) der Melodiestimme zu erfassen [SG12, Kla08]. Auf Basis dieser F0-Schätzung kann dann die Stimme vom Klanggemisch abgetrennt werden. Im Folgenden skizzieren wir ein solches Verfahren anhand der Abbildungen 7 und 8.

In unserem Beispielszenario soll die Gesangsstimme aus der Aufnahme eines Klavierlieds (Gesang begleitet von Klavier) extrahiert werden. In einem ersten Schritt wird der Verlauf der Grundfrequenz der Singstimme wie in [SG12] ermittelt. Ausgehend von einem Spektrogramm mit linearer Frequenzachsenaufteilung wird zunächst ein sogenanntes *Log-Spektrogramm* mit logarithmischer Frequenzachsenaufteilung berechnet. Diese Darstellung hat den Vorteil, dass so der Abstand eines Obertones zu seinem Grundton unabhängig von der jeweiligen Frequenz des Grundtons ist, siehe auch Abbildung 7b. Zur Berechnung des Log-Spektrogramms wird zunächst die Frequenzauflösung des Spektro-

[2] Diese werden in Abbildung 5c und Abbildung 6a durch rote Umrandungen wiedergegeben.

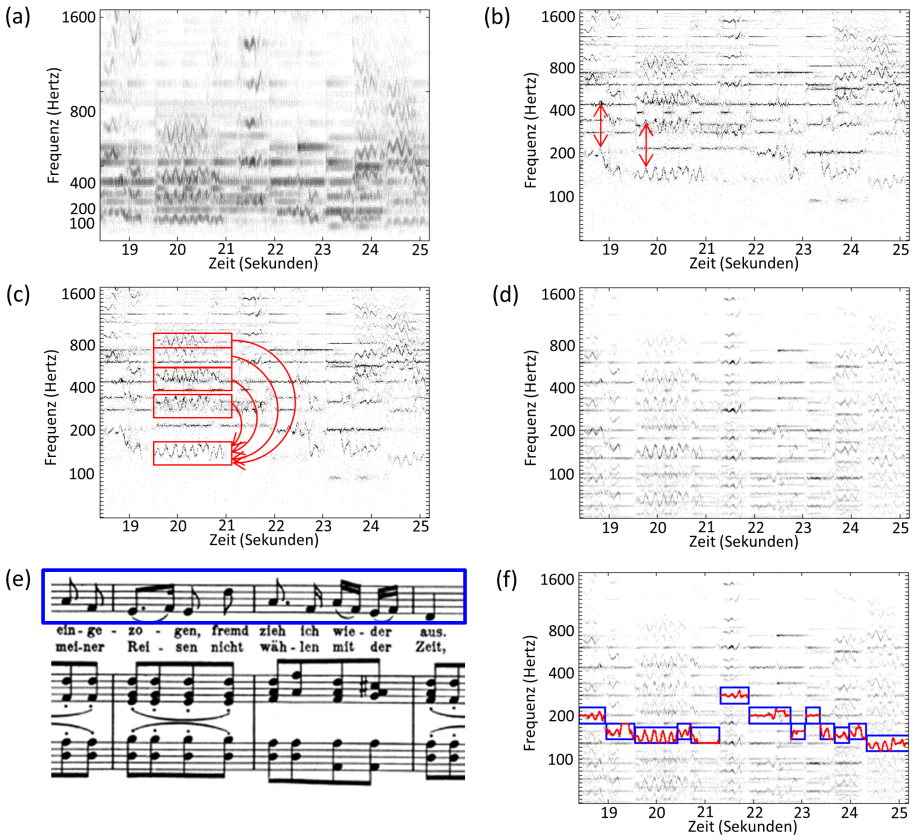

Abbildung 7: Notentext-informierte F0-Schätzung. **(a)** Spektrogramm eines Ausschnittes von "Gute Nacht" aus Schumanns "Winterreise". **(b)** Log-Spektrogramm. **(c)** Berechnung der Salienz-Darstellung. **(d)** Salienz-Darstellung. **(e)** Notentext mit Singstimme. **(f)** F0-Schätzung der Gesangsstimme.

gramms unter Zuhilfenahme der Phaseninformation verfeinert, und dann die Frequenzen logarithmisch aufgeteilten Frequenzbänder zugeteilt. Da die Singstimme typischer Weise energiereiche Obertöne aufweist, werden zur Hervorhebung des Grundfrequenzverlaufs zu jedem Eintrag im Log-Spektrogramm die Einträge der zugehörigen Obertöne addiert, siehe Abbildung 7c. Die so erhaltene *Salienz-Darstellung* gibt zu einem gegebenen Zeitpunkt an, wie stark ein Ton einer bestimmten Grundfrequenz aus dem Klanggemisch hervorsticht. In [SG12] wird nun unter Annahme von zeitlichen und spektralen Kontinuitätseigenschaften versucht, den Verlauf der Melodielinie auf Basis der Salienz-Darstellung zu bestimmen. Als Alternative kann auch ein Notentext-informierter Ansatz verfolgt werden, bei dem mit Hilfe einer synchronen Klavierwalzendarstellung die Salienz-Darstellung auf den Bereich des groben Grundfrequenzverlaufs einschränkt wird. Durch Verwendung eines auf dynamischer Programmierung basierenden Algorithmus

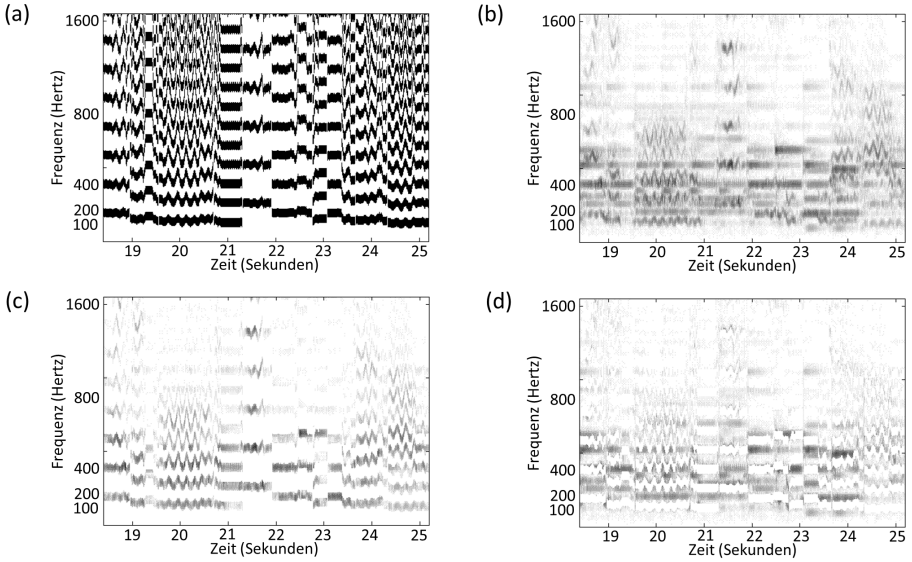

Abbildung 8: Abtrennung der Gesangsstimme mittels Maskierung. **(a)** F0-basierte binäre Maskierungsmaske für die Gesangsstimme. **(b)** Spektrogramm der Originalaufnahme (Gesang und Klavier). **(c)** Spektrogramm der Gesangsstimme. **(d)** Spektrogramm des Restsignals.

kann dann die Frequenztrajektorie der Singstimme innerhalb dieser Bereiche effizient berechnet werden, siehe Abbildung 7f. Der errechnete Grundfrequenzverlauf erlaubt es nun, eine binäre Maske für das ursprüngliche Spektrogramm zu erstellen. Hierbei bleiben nur diejenigen Spektrogrammeinträge, die zum Frequenzverlauf der Singstimme und deren Obertönen korrespondieren, erhalten (Abbildung 8a). Punktweise Multiplikation der Maske mit dem Spektrogramm resultiert im Spektrogramm der Gesangsstimme. Durch Anwendung einer inversen gefensterten Fouriertransformation erhält man schließlich die zugehörige Wellenform. Zusätzlich kann noch der verbleibende Teil des Spektrogramms betrachtet werden durch den man in dem von uns betrachteten Szenario die Klavierstimme der Aufnahme erhält (Abbildung 8d).

7 Anwendungen und Fazit

In diesem Artikel haben wir gezeigt, wie sich eine Audioaufnahme unter Ausnutzung von Notentextinformation in elementare Bausteine zerlegen lässt. Solche Zerlegungen sind nicht nur für das Verständnis der zugrundeliegenden Musiksignale von grundlegender Bedeutung, sondern ermöglichen auch eine Reihe von neuartigen Anwendungen. Zum Abschluss dieses Artikels skizzieren wir exemplarisch zwei solche Anwendungen, die das Potential der notentext-informierten Quellentrennung andeuten sollen.

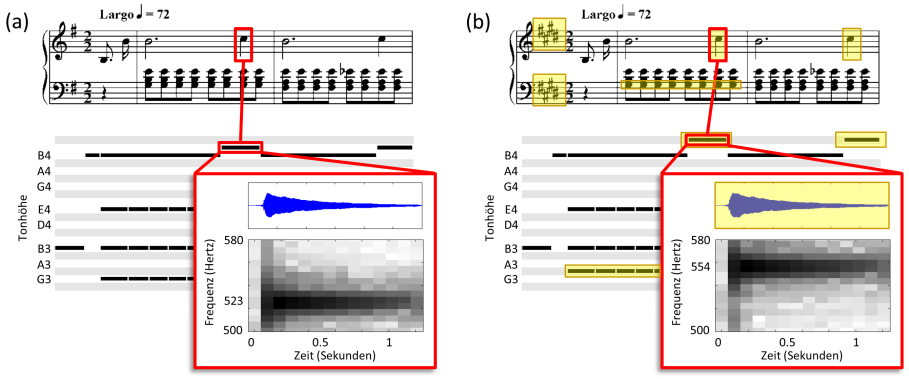

Abbildung 9: Notentext-informierte Editierung von Audiomaterial, bei der eine Audioaufnahme von Moll nach Dur moduliert wird. Die manipulierten Elemente (rechts) sind gelb unterlegt.

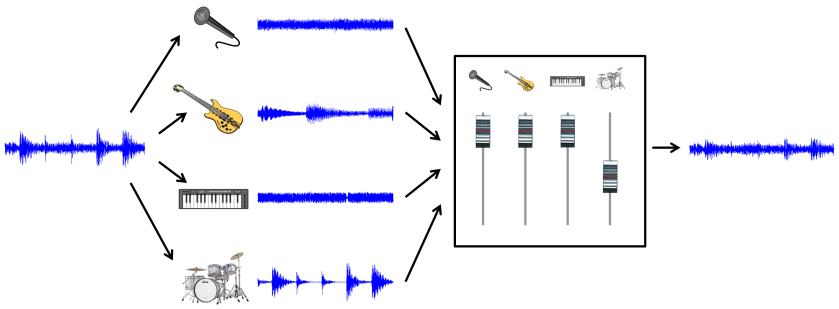

Abbildung 10: Intelligenter Equalizer zur Veränderung von Sing- und Instrumentalstimmen.

In Abbildung 9 ist eine Benutzerschnittstelle zur Notentext-basierten Editierung von Audiodaten angedeutet. Mittels der notenbasierten Zerlegung können Manipulation im Notentext auf das Musiksignal übertragen werden. Zum Beispiel kann auf diese Weise, wie in Abbildung 9b illustriert, die Audioaufnahme von Moll nach Dur moduliert werden. Eine zweite Anwendung ist in Abbildung 10 skizziert. Auf Basis einer Zerlegung in Einzelstimmmen können *intelligente Equalizer* realisiert werden, bei denen ein Benutzer anstelle fester Frequenzbänder semantisch sinnvolle Einheiten wie Instrumental- und Singstimmen verstärken oder abschwächen kann [IGK+08].

Ziel dieses Artikels war es, über neuartige Entwicklungen im Bereich der Quellentrennung von Musiksignalen zu berichten. Trotz erheblicher Forschungsbemühungen steckt man bei diesen extrem schwierigen Fragestellungen noch in den Kinderschuhen. Selbst bei Notentext-informierten Verfahren weisen die abgetrennten Quellen oft noch starke Artefakte auf, so dass hier noch erheblicher Forschungsbedarf besteht bis diese Techniken anwendungstauglich werden.

Literatur

[Bre90] Albert S. Bregman. *Auditory scene analysis: the perceptual organization of sound*. MIT Press, 1990.

[BW05] Mark A. Bartsch und Gregory H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Trans. on Multimedia*, 7(1):96–104, 2005.

[Che53] Collin E. Cherry. Some Experiments on the Recognition of Speech, with One and with Two Ears. *Journal of the Acoustic Society of America (JASA)*, 24:975–979, 1953.

[Dre11] Karin Dressler. An auditory streaming approach for melody extraction from polyphonic music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Seiten 19–24, Miami, USA, 2011.

[EM11] Sebastian Ewert und Meinard Müller. Estimating note intensities in music recordings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seiten 385–388, Prague, Czech Republic, 2011.

[EM12a] Sebastian Ewert und Meinard Müller. Score-informed source separation for music signals. In Meinard Müller, Masataka Goto und Markus Schedl, Hrsg., *Multimodal Music Processing*, Jgg. 3 of *Dagstuhl Follow-Ups*, Seiten 73–94. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2012.

[EM12b] Sebastian Ewert und Meinard Müller. Using score-informed constraints for NMF-based source separation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seiten 129–132, Kyoto, Japan, 2012.

[EMG09] Sebastian Ewert, Meinard Müller und Peter Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seiten 1869–1872, Taipei, Taiwan, 2009.

[Fit10] Derry Fitzgerald. Harmonic/percussive separation using median filtering. In *Proc. International Conference on Digital Audio Effects (DAFX)*, Graz, Austria, 2010.

[Got00] Masataka Goto. A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, Jgg. 2, Seiten 757–760, 2000.

[Got04] Masataka Goto. A Real-time Music-scene-description System: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication (ISCA Journal)*, 43(4):311–329, 2004.

[GR08] Olivier Gillet und Gaël Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3):529–540, 2008.

[HBD10] Romain Hennequin, Roland Badeau und Bertrand David. Time-dependent parametric and harmonic templates in non-negative matrix factorization. In *Proc. International Conference on Digital Audio Effects (DAFx)*, Seiten 246–253, Graz, Austria, 2010.

[HDB11] Romain Hennequin, Bertrand David und Roland Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seiten 45–48, Prague, Czech Republic, 2011.

[HKV09] Toni Heittola, Anssi P. Klapuri und Tuomas Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, Seiten 327–332, Kobe, Japan, 2009.

[HR07] Yushen Han und Christopher Raphael. Desoloing monaural audio using mixture models. In *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, Seiten 145–148, Vienna, Austria, 2007.

[IGK+08] Katsutoshi Itoyama, Masataka Goto, Kazunori Komatani, Tetsuya Ogata und Hiroshi G. Okuno. Instrument equalizer for query-by-example retrieval: Improving sound source separation based on integrated harmonic and inharmonic models. In *Proc. International Conference for Music Information Retrieval (ISMIR)*, Seiten 133–138, Philadelphia, USA, 2008.

[JER11] Cyril Joder, Slim Essid und Gaël Richard. A conditional random field framework for robust and scalable audio-to-score matching. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2385–2397, 2011.

[Kla08] Anssi P. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):255–266, 2008.

[LS00] Daniel D. Lee und H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proc. Neural Information Processing Systems (NIPS)*, Seiten 556–562, Denver, CO, USA, 2000.

[MEKR11] Meinard Müller, Daniel P. W. Ellis, Anssi Klapuri und Gaël Richard. Signal processing for music analysis. *IEEE Journal on Selected Topics in Signal Processing*, 5(6):1088–1110, 2011.

[Mül07] Meinard Müller. *Information retrieval for music and motion*. Springer Verlag, 2007.

[OMKS08] Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka und Shigeki Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, Seiten 139–144, Philadelphia, Pennsylvania, USA, 2008.

[PEE+07] Graham E. Poliner, Daniel P.W. Ellis, Andreas F. Ehmann, Emilia Gómez, Sebastian Streich und Beesuan Ong. Melody transcription from music audio: approaches and evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1247–1256, 2007.

[RK08] Matti Ryynänen und Anssi P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008.

[ROS07] Stanislaw Andrzej Raczynski, Nobutaka Ono und Shigeki Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, Seiten 381–386, 2007.

[RP09] Lise Regnier und Geoffroy Peeters. Singing voice detection in music tracks using direct voice vibrato detection. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seiten 1685–1688, Taipei, Taiwan, 2009.

[SG12] Justin Salamon und Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech & Language Processing*, 20(6):1759–770, 2012.

[SRS08] Madhusudana Shashanka, Bhiksha Raj und Paris Smaragdis. Probabilistic latent variable models as nonnegative factorizations (Article ID 947438). *Computational Intelligence and Neuroscience*, 2008.

[WPD06] John Woodruff, Bryan Pardo und Roger B. Dannenberg. Remixing stereo music with score-informed source separation. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, Seiten 314–319, 2006.

[WVR+11] Jun Wu, Emmanuel Vincent, Stanislaw Andrzej Raczynski, Takuya Nishimoto, Nobutaka Ono und Shigeki Sagayama. Multipitch estimation by joint modeling of harmonic and transient sounds. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seiten 25–28, Prague, Czech Republic, 2011.