Harmonisch-Perkussiv-Rest Zerlegung von Musiksignalen

Jonathan Driedger, Meinard Müller
International Audio Laboratories Erlangen, D-91058 Erlangen,
E-Mail: {jonathan.driedger, meinard.mueller}@audiolabs-erlangen.de

Einleitung

Die Zerlegung von komplexen Musiksignalen in semantisch interpretierbare Einzelbestandteile ist eine zentrale Fragestellung der Musiksignalverarbeitung. Einige dieser Bestandteile sind von harmonischer Natur, die sich als horizontale Strukturen in Zeit-Frequenz-Darstellungen niederschlagen. Andere Bestandteile sind eher perkussiv und führen zu vertikalen Strukturen. Basierend auf diesen Beobachtungen wurden in der Vergangenheit unterschiedliche Verfahren zur harmonisch-perkussiven Zerlegung (*HP-Zerlegung*) von Audiosignalen entwickelt. Ein Problem dieser Verfahren besteht darin, dass alle Bestandteile des Audiosignals entweder der harmonischen oder der perkussiven Komponente zugeordnet werden. Dies ist allerdings nicht immer sinnvoll, da es auch rauschartige oder fluktuierende Bestandteile geben kann, die weder harmonisch noch perkussiv sind. In diesem Beitrag berichten wir von einem neuartigen Verfahren, bei dem ein Signal in drei Komponenten zerlegt wird: eine klar harmonische, eine klar perkussive, sowie eine dazwischenliegende Restkomponente (*HPR-Zerlegung*). Die Zerlegung lässt sich hierbei durch einen Trennungsparameter beeinflussen. Exemplarisch zeigen wir, wie sich diese flexible Zerlegung für die Verbesserung gängiger Audiomerkmale zur Erfassung tonaler oder rhythmischer Signalcharakteristika einsetzen lässt. Eine weitere vielversprechende Anwendung besteht im Bereich der Quellentrennung, bei der zum Beispiel die Singstimme aus einer polyphonen Musikaufnahme herausgetrennt werden soll. Hier enthält die Restkomponente oft die für den Gesang sehr typischen Bestandteile die zu Vibrato und Glissando korrespondieren.

Harmonisch-Perkussiv-Rest Zerlegung

Viele Verfahren zur Zerlegung eines Musiksignals in eine harmonische und eine perkussive Komponente (HP-Zerlegung) basieren auf der Beobachtung, dass sich in einer Spektrogramm-Darstellung des Signals harmonische Klänge als horizontale und perkussive Klänge als vertikale Strukturen manifestieren ([5, 10]). Abhängig davon, ob ein Zeit-Frequenz-Punkt im Spektrogramm einer Musikaufnahme eher zu einer horizontalen oder einer vertikalen Struktur gehört, wird dieser Punkt entweder der harmonischen oder der perkussiven Komponente zugeordnet. Das Spektrogramm wird so in zwei Teilspektrogramme zerlegt, die anschließend wieder in Zeitsignale überführt werden. Diese Signale bilden die harmonische und die perkussive Komponente des Eingangssignals. Rauschartige oder fluktuierende Klänge, die weder horizontale noch vertikale Strukturen im Spektrogramm hervorrufen, werden bei dieser Methode mehr oder weniger zufällig zwischen der harmonischen und der perkussiven Komponente aufgeteilt. Beispiele für solche Klänge sind etwa rauschartiger Applaus oder eine durch Vibrato und Glissando ausgeschmückte Singstimme.

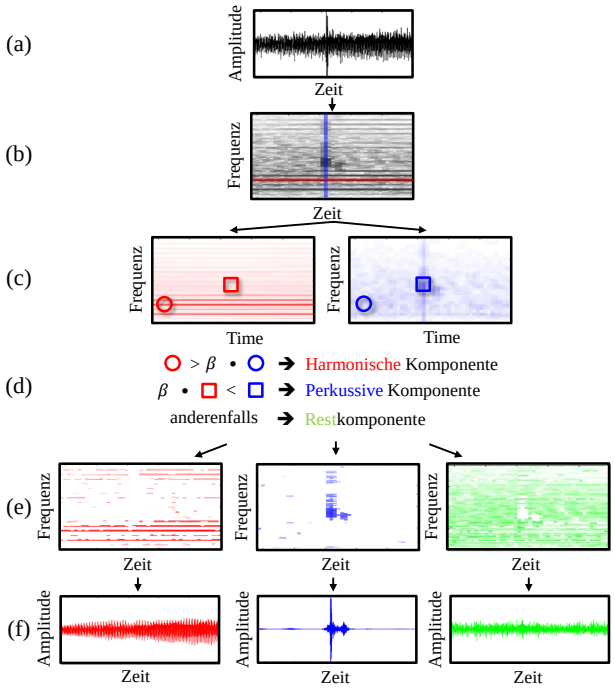

Abbildung 1: Harmonisch-Perkussiv-Rest Zerlegung- **(a):** Musiksignal. **(b):** Spektrogramm mit angedeuteten harmonischen und perkussiven Strukturen. **(c):** Gefilterte Spektrogramme. **(d):** Zerlegungsvorschrift. **(e):** Spektrogramme der harmonischen, der perkussiven und der Restkomponente. **(f):** Rekonstruierte Audiosignale der harmonischen, der perkussiven und der Restkomponente.

Kürzlich wurde von uns in [3] ein Verfahren vorgestellt, welches die klassische HP-Zerlegung um eine dritte Komponente erweitert. Ziel dieses neuartigen Verfahrens ist es, ein gegebenes Musiksignal in eine klar harmonische, eine klar perkussive und eine *Restkomponente* zu zerlegen (HPR-Zerlegung). Die Restkomponente enthält dabei Klänge, die weder klar harmonisch noch klar perkussiv sind. Ein gegebenes Musiksignal (Abbildung 1a) wird zunächst mit Hilfe der gefensterten Fouriertransformation in ein Spektrogramm überführt (Abbildung 1b). Wie in [5] werden nun horizontale bzw. vertikale Strukturen durch die Anwendung eines horizontalen bzw. vertikalen Median-Filters verstärkt (Abbildung 1c). Basierend auf den zwei gefilterten Spektrogrammen wird im Anschluss jeder Zeit-Frequenz Wert des ursprünglichen Spektrogramms entweder der harmonischen, der perkussiven, oder der Restkomponente zugeordnet (Abbildung 1d). Zu

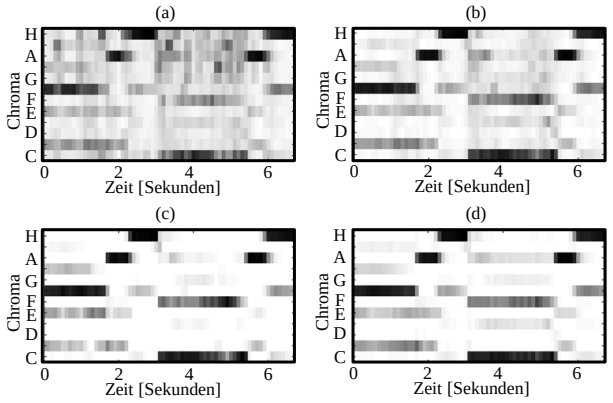

Abbildung 2: Chroma-Darstellung verschiedener Signalkomponenten. **(a):** Gesamtes Klanggemisch bestehend aus Geige, Kastagnetten und Applaus. **(b):** Harmonische Komponente der klassischen HP-Zerlegung. **(c):** Harmonische Komponente der vorgestellten HPR-Zerlegung mit $\beta = 2$. **(d):** Isolierte Geigenstimme.

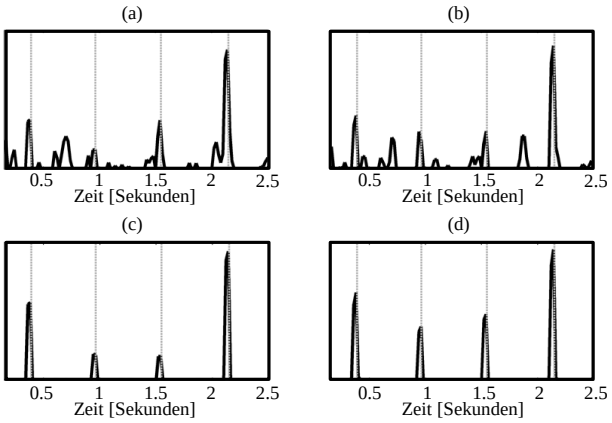

Abbildung 3: Novelty-Kurven verschiedener Signalkomponenten. **(a):** Gesamtes Klanggemisch bestehend aus Geige, Kastagnetten und Applaus. **(b):** Perkussive Komponente der klassischen HP-Zerlegung. **(c):** Perkussive Komponente der vorgestellten HPR-Zerlegung mit $\beta = 3$. **(d):** Isolierte Kastagnetten.

diesem Zweck führen wir einen *Trennungsparameter* β ein. Ein Zeit-Frequenz Wert des ursprünglichen Spektrogramms wird genau dann der harmonischen Komponente zugeordnet, wenn der Wert im horizontal gefilterten Spektrogramm den entsprechenden Wert im vertikal gefilterten Spektrogramm um einen Faktor von β übertrifft. Umgekehrt wird ein Zeit-Frequenz Wert der perkussiven Komponente zugeordnet, wenn der Wert im vertikal gefilterten Spektrogramm um den Faktor β größer ist als der entsprechende Wert im horizontal gefilterten Spektrogramm. Zeit-Frequenz Werte, die keine dieser Bedingungen erfüllen, werden der Restkomponente zugeordnet. Die drei resultierenden Spektrogramme (Abbildung 1e) werden anschließend mit Hilfe einer inversen gefensterten Fouriertransformation [7] und unter Verwendung der Phaseninformationen des ursprünglichen Klanggemischs in den Zeitbereich überführt. Die sich ergebenden Zeitsignale bilden die drei gewünschten Signalkomponenten (Abbildung 1f).

Im folgenden diskutieren wir einige mögliche Anwendungen der vorgestellten HPR-Zerlegung.

Verbesserung von Audiomerkmalen

Ein möglicher Anwendungsbereich des vorgestellten Zerlegungsverfahrens ist die Verbesserung von Audiomerkmalen. Die klassische HP-Zerlegung wurde in diesem Kontext bereits erfolgreich angewandt, so zum Beispiel in [11] zur Verbesserung von Chroma-Merkmalen oder in [6] im Zusammenhang mit Tempo-Merkmalen. Die HPR-Zerlegung bietet die Möglichkeit, solche Merkmale noch weiter zu verbessern. Um dieses Potential anzudeuten, betrachten wir exemplarisch ein künstliches Klanggemisch aus einer Geige (klar harmonisch), Kastagnetten (klar perkussiv) und Applaus (rauschartig).

In Abbildung 2a sehen wir die Chroma-Merkmale des betrachteten Signals, berechnet mit der Chroma-Toolbox [9]. Diese Merkmale erfassen die zeitabhängige Energieverteilung des Signals in Bezug auf die in der westlichen Musik üblichen 12 Tonklassen $C, C^{\#}, D, \ldots, H$. Die Folge solcher Chroma-Merkmale korreliert zum harmonischen Verlauf der Musikaufnahme. In Anwendungen ist es häufig von Vorteil, wenn die berechneten Merkmale nicht durch perkussive oder rauschartige Klänge in der Musikaufnahme verunreinigt sind. Solche Verunreinigungen sind zum Beispiel in der Chroma-Darstellung unseres Beispiels zu sehen (Abbildung 2a). Die durch die Geige getragene harmonische Information wird durch die Kastagnetten und den Applaus stark verrauscht. Um dieses Rauschen zu verringern, besteht ein Ansatz darin, die Merkmale lediglich auf der harmonischen Komponente der HP-Zerlegung zu berechnen [11]. Die resultierende Chroma-Darstellung ist etwas weniger verrauscht (Abbildung 2b). Trotzdem schlägt sich der in dieser Komponente immer noch hörbare Applaus auch in den Chroma-Merkmalen nieder. Die Verwendung der vorgestellten HPR-Zerlegung erlaubt es, rauschartige Klänge in der harmonischen Komponente zu unterdrücken. In Abbildung 2c sieht man die Chroma-Merkmale der harmonischen Komponente dieser Zerlegung (berechnet mit $\beta = 2$). Die gezeigten Merkmale sind kaum verunreinigt und ähneln stark den Chroma-Merkmalen der isolierten Geigenstimme (Abbildung 2d). Da die Geige in dem betrachteten Klanggemisch aus Geige, Kastagnetten und Applaus die einzige harmonische Stimme ist, kann ihre Chroma-Darstellung als Referenz angenommen werden.

Ein ähnlicher Ansatz kann zur Verbesserung von Novelty-Kurven verwendet werden, die häufig zur Erkennung von Noteneinsatzzeiten (Onsets) in Musiksignalen Anwendung finden, siehe zum Beispiel [8]. Peaks in der Novelty-Kurve weisen dabei auf potentielle Onsets in dem betrachteten Musiksignal hin. In Abbildung 3a sehen wir einen Auschnitt der Novelty-Kurve unseres vorherigen Beispielsignals. Die zeitlichen Positionen der in diesem Signal hörbaren Onsets, verursacht durch die Kastagnetten, sind durch graue Linien im Hintergrund angedeutet. Zu jedem Onset im Signal tritt auch ein Peak in der Novelty-Kurve auf. Durch den Applaus in der Aufnahme entstehen jedoch weitere Peaks, die eine automatisierte Extraktion der Onset-Positionen der Kastagnetten erschweren. Die Novelty-Kurve der perkussiven Kom-

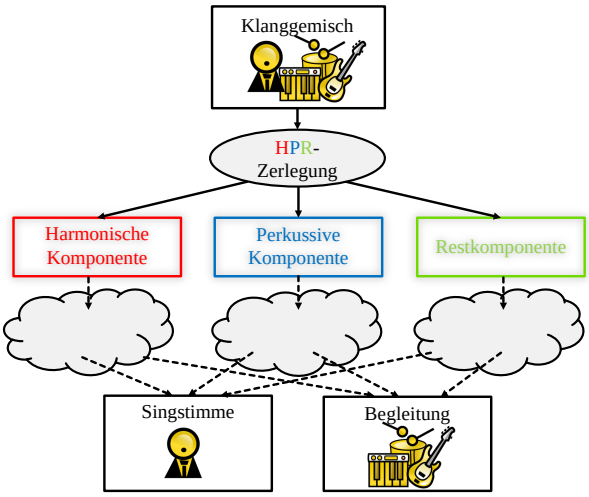

Abbildung 4: HPR-Zerlegung zur Singstimmenabtrennung.

ponente der klassischen HP-Zerlegung (Abbildung 3b) ähnelt der Novelty-Kurve des Klanggemisches. Der Applaus ist in dieser Komponente noch zu großen Teilen enthalten und verursacht auch hier unerwünschte Peaks. Durch die Verwendung der HPR-Zerlegung mit $\beta = 3$ können diese Applauskomponenten fast komplett aus der perkussiven Komponente entfernt werden. Die resultierende Novelty-Kurve (Abbildung 3c) weist keine Verunreinigungen mehr auf. Auch hier ähnelt die berechnete Kurve stark der Novelty-Kurve des als Referenz dienenden isolierten Kastagnettensignals (Abbildung 3d).

Diese beiden Beispiele deuten an, wie sich die HPR-Zerlegung bei der Berechnung von Audiomerkmalen gewinnbringend verwenden lassen könnte.

Singstimmentrennung in polyphonen Musikaufnahmen

Ein weiteres mögliches Anwendungsgebiet der HPR-Zerlegung ist die automatisierte Abtrennung der Gesangsstimme aus einer mehrstimmigen Musikaufnahme. Die menschliche Singstimme ist in der Lage, eine Vielzahl von verschiedenen Klängen zu erzeugen. Neben klaren harmonischen Komponenten sind auch stimmlose Laute, wie etwa Frikative oder Atemgeräusche, wichtig für den Klang einer Stimme. Weiterhin verwenden Sänger häufig Vibrato oder Glissando, um ihrem Gesang mehr Ausdruck zu verleihen. Diese verschiedenen Klänge führen zu sehr unterschiedlichen spektralen Strukturen. Der tonale Anteil der Stimme weist typischerweise klare Obertonverläufe auf, während Frikative sich in vertikalen Strukturen manifestieren und Atemgeräusche rauschartig sind. Diese Strukturvielfalt macht es schwer, die gesamte Singstimme mit nur einem Modell zu erfassen. Um in einer Musikaufnahme Klänge, die zu der Singstimme gehören, von Begleitstimmen zu unterscheiden, sind solche expliziten Modellannahmen jedoch häufig notwendig. In einer kürzlich vorgestellten Arbeit haben wir gezeigt, wie sich dieses Problem mit Hilfe der HPR-Zerlegung in einfacher zu lösende Teilprobleme zerlegen lässt [2]. Der vorgestellte Ansatz wird in Abbildung 4 skizziert. Zerlegt man eine Musikaufnahme mit Hilfe der HPR-Zerlegung in eine harmonische, eine perkussive und eine Restkomponente, so sind in allen drei Komponenten sowohl Anteile der Singstimme, als auch Anteile der Begleitstimmen zu hören. Die harmonische Komponente enthält beispielsweise sowohl die klar harmonischen Anteile der Singstimme, als auch die der begleitenden harmonischen Instrumente. Da diese Komponente aber keine perkussiven oder rauschartigen Klänge mehr enthält (z.B. Frikative, Transienten oder Atemgeräusche) lassen sich in einem weiteren Zerlegungsschritt die Klänge der Singstimme wesentlich einfacher von denen der Begleitstimmen trennen. Das Gleiche gilt für die perkussive und die Restkomponente. In einem letzten Schritt werden die einzelnen Klangkomponenten der Singstimme und der Begleitung wieder zusammengefügt.

Auf der unter [1] publizierten Webseite sind viele Beispiele für Trennungsergebnisse zu finden. Hier kann man sich auch die einzelnen Zwischenstufen der in dem Trennungsverfahren berechneten Signale anhören.

Zusammenfassung

In diesem Betrag wurden Anwendungsbeispiele eines kürzlich vorgestellten Verfahrens zur Zerlegung eines Musiksignales in eine harmonische, eine perkussive und eine Restkomponente skizziert. Eine Idee war, ein Audiomerkmal lediglich auf der für das Merkmal maßgeblichen Klangkomponente zu berechnen, was zu „reineren“ Merkmalsdarstellungen führen kann. Dieser Ansatz wurde am Beispiel von Chroma-Merkmalen und Novelty-Kurven verdeutlicht. Ein weiterer, möglicher Anwendungsbereich ist die Abtrennung der Singstimme aus einer gegebenen Musikaufnahme. Bei dieser Aufgabenstellung stellt die Unterscheidung der komplexen Klängen der Singstimme von den Klängen der musikalischen Begleitung ein Hauptproblem dar. Hier hat sich gezeigt, dass dieses Problem durch die Anwendung der HPR-Zerlegung in einfachere Teilprobleme untergliedert werden kann. Ähnliche Ansätze könnten sich auch auf andere Aufgabenbereiche in der Signalverarbeitung übertragen lassen. In zukünftiger Arbeit wollen wir etwa untersuchen, ob sich die HPR-Zerlegung, ähnlich wie in [4], auch für die Verbesserung von Algorithmen zur Zeitstreckung von Musiksignalen (Time-Scale Modification) verwenden lässt.

Literatur

[1] Jonathan Driedger and Meinard Müller. Accompanying website: Extracting singing voice from music recordings by cascading audio decomposition techniques. `http://www.audiolabs-erlangen.de/resources/MIR/2015-ICASSP-SVECD`.

[2] Jonathan Driedger and Meinard Müller. Extracting singing voice from music recordings by cascading audio decomposition techniques. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[3] Jonathan Driedger, Meinard Müller, and Sascha Disch. Extending harmonic-percussive separation of audio signals. In *Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR)*, pages 611–616, Taipei, Taiwan, 2014.

[4] Jonathan Driedger, Meinard Müller, and Sebastian Ewert. Improving time-scale modification of music signals using harmonic-percussive separation. *Signal Processing Letters, IEEE*, 21(1):105–109, 2014.

[5] Derry Fitzgerald. Harmonic/percussive separation using median filtering. In *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, pages 246–253, Graz, Austria, 2010.

[6] Aggelos Gkiokas, Vassilios Katsouros, George Carayannis, and Themos Stafylakis. Music tempo estimation and beat tracking by applying source separation and metrical relations. In *ICASSP*, pages 421–424, 2012.

[7] Daniel W. Griffin and Jae S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2):236–243, 1984.

[8] Peter Grosche and Meinard Müller. Extracting predominant local pulse information from music recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1688–1701, 2011.

[9] Meinard Müller and Sebastian Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 215–220, Miami, FL, USA, 2011.

[10] Nobutaka Ono, Kenichi Miyamoto, Hirokazu Kameoka, and Shigeki Sagayama. A real-time equalizer of harmonic and percussive components in music signals. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 139–144, Philadelphia, Pennsylvania, USA, 2008.

[11] Yushi Ueda, Yuuki Uchiyama, Takuya Nishimoto, Nobutaka Ono, and Shigeki Sagayama. HMM-based approach for automatic chord detection using refined acoustic features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5518–5521, Dallas, USA, 2010.