Markus Schwabe*, Marian Weber und Fernando Puente León

Notenseparation in polyphonen Musiksignalen durch einen Matching-Pursuit-Algorithmus

Note separation in polyphonic music signals with a matching pursuit algorithm

DOI 10.1515/teme-2018-0039

Zusammenfassung: Aus einem polyphonen Musiksignal mit mehreren gleichzeitig klingenden Instrumenten werden die zu einzelnen Noten gehörenden Töne mithilfe eines harmonischen Matching-Pursuit-Algorithmus (HMPA) separiert. Hierfür wird im ersten Schritt des Verfahrens die für den Matching-Pursuit-Algorithmus notwendige Bibliothek auf Basis des vorliegenden Musiksignals aufgebaut. Im zweiten Schritt werden die in der Bibliothek hinterlegten Notenmodelle zur Approximation des Musiksignals verwendet. Anschließend lassen sich die einzelnen Noten aus der Approximation extrahieren und mit dem Vorwissen von Melodieverläufen zu separierten Melodien einzelner Instrumente zusammenfassen.

Schlüsselwörter: Notenseparation, harmonischer Matching-Pursuit-Algorithmus, blinde Quellentrennung, polyphone Musiksignale, Signalrekonstruktion.

Abstract: Sound signals that correspond to discrete notes are separated from a polyphonic music signal including several instruments playing at the same time with a harmonic matching pursuit algorithm (HMPA). The required dictionary for the matching pursuit algorithm is developed in the first step of this method on the basis of the current music signal. In the second step, the note models that are stored in the dictionary are employed to approximate the music signal. The discrete notes can be extracted from this approximation and can be combined to separated melodies of occurrent instruments by means of previous knowledge about melodic lines.

Keywords: Note separation, harmonic matching pursuit algorithm, blind source separation, polyphonic music signals, signal reconstruction.

***Korrespondenzautor: Markus Schwabe,** Institut für Industrielle Informationstechnik (IIIT), Karlsruher Institut für Technologie (KIT), Hertzstraße 16, 76187 Karlsruhe, E-Mail: markus.schwabe@kit.edu

Marian Weber, Fernando Puente León, Institut für Industrielle Informationstechnik (IIIT), Karlsruher Institut für Technologie (KIT), Hertzstraße 16, 76187 Karlsruhe, E-Mail: marian.weber@student.kit.edu, puente@kit.edu

1 Einleitung

In der digitalen Bearbeitung von Musiksignalen können einzelne Töne oder Melodien in ihrer Tonhöhe, ihrem Klang, ihrer Dauer und ihrer Geschwindigkeit bearbeitet werden, wenn sie separat vorliegen. Dadurch können beispielsweise einzelne Fehler in Musikaufnahmen korrigiert oder gewünschte Passagen aus einem Musikstück in ein anderes integriert werden. Bei mehrstimmigen Musiksignalen, welche auch als polyphon bezeichnet werden, treten gleichzeitig mehrere Töne und Melodien auf, die für eine derartige Nachbearbeitung getrennt werden müssen.

Eine weit verbreitete Methode zur blinden Quellentrennung, die auch in der Sprach- und Musikseparation große Anwendung findet, ist die *Nonnegative Matrix Factorisation* (NMF). Für die Notenseparation wurde sie das erste Mal von Smaragdis und Brown [16] umgesetzt. Dabei wird die Zeit-Frequenz-Darstellung des Musiksignals als Matrix aufgefasst und mithilfe einer Optimierung in die Multiplikation von zwei kleineren, nichtnegativen Matrizen zerlegt. Damit stellt die NMF einen unüberwachten Ansatz dar, benötigt aber für eine genaue Approximation die Anzahl der im Musiksignal auftretenden Noten und Instrumente sowie möglichst jede Note mindestens einmal monophon aufgenommen [13].

Neben dem unüberwachten Ansatz haben sich auch halb oder komplett überwachte Methoden durchgesetzt, bei denen Vorwissen über die vorkommenden Instrumente durch Lernen der charakteristischen Instrumentenspektren genutzt wird. Die bisher besten Ergebnisse für die Transkription von Noten wurden durch überwachte Verfahren mit neuronalen Netzen (NN) erzielt. Mithilfe von umfangreichen Datensätzen werden die NN trainiert und damit auf die Erkennung von Noten der gelernten Instrumente spezialisiert. Dabei können unterschiedliche Netzarchitekturen wie *Convolutional Neural Networks* (CNN) [15] oder *Recurrent Neural Networks* (RNN) [8] gewählt werden.

Der dritte Ansatz, der wie die NMF und NN in vielen Arbeiten zur Notenseparation und -transkription eingesetzt wird, ist der Matching-Pursuit-Algorithmus (MPA).

Dieser beinhaltet in einer vordefinierten Bibliothek Signalmodelle unterschiedlicher Parameter, sogenannte Atome, und approximiert das zu analysierende Signal anhand der ähnlichsten Atome. Um das Musiksignal mit spezifischen Notenmodellen zu approximieren, führten Gribonval und Bacry den harmonischen MPA (HMPA) mit harmonischen Atomen ein [6]. Eine sehr performante Implementierung des HMPA stellt der Algorithmus des *Matching Pursuit Tool Kit* (MPTK) [9] dar. Dessen Qualität der Notenseparation hängt allerdings stark von der Wahl geeigneter Notenspektren der Atome in der Bibliothek ab. Aus diesem Grund führten Carabias-Orti et al. [2] vor dem HMPA eine Schätzung der vorkommenden Notenspektren im Frequenzbereich mithilfe eines Expectation-Maximization-Algorithmus durch. Auf Basis dieser Schätzung wird eine an das Signal angepasste Bibliothek erstellt, die dem MPTK-Algorithmus zugrunde gelegt wird. Diese signaladaptive Notenseparation nach Carabias-Orti et al. ist nur für Musiksignale eines Instruments entwickelt [2].

Insgesamt zeichnet sich der MPA gegenüber den anderen beiden Verfahren durch eine sehr feine Zeit- und Frequenzauflösung sowie der Unabhängigkeit von umfangreichen Lerndatensätzen oder spezifischem Vorwissen aus. Deshalb wird er im vorliegenden Ansatz zur Notenseparation monauraler polyphoner Musiksignale umgesetzt. Um eine möglichst zum Musiksignal passende Bibliothek aufbauen zu können, werden die relevanten Notenspektren aller im Signal enthaltenen Instrumente vor der Separation geschätzt und daraus die relevanten Atome gebildet.

In Abschnitt 2 wird der Matching-Pursuit-Algorithmus und seine Erweiterung für Musiknoten vorgestellt. Anschließend erläutert Abschnitt 3 die umgesetzten Weiterentwicklungen des HMPA zur Notenseparation polyphoner Musiksignale mit mehreren Instrumenten. Die damit erzielten Ergebnisse werden in Abschnitt 4 diskutiert.

2 Matching-Pursuit-Algorithmus

Zunächst wird die allgemeine Funktionsweise des Matching-Pursuit-Algorithmus beschrieben. Anschließend werden die für den harmonischen MPA (HMPA) notwendigen Atome definiert und eine schnelle Umsetzung des HMPA vorgestellt, die in dieser Arbeit verwendet wird.

2.1 Definition

Der MPA nach Mallat und Zhang [11] ist eine Methode, die auf dem Ähnlichkeitsvergleich eines Signals mit vielen Zeit-Frequenz-Atomen, welche in einem übervollständigen Wörterbuch $\mathcal{D}$ zusammengefasst sind, basiert. Die Ähnlichkeit wird dabei durch die Korrelation C des Signals $s(t)$ und des Atoms $g(t)$ berechnet. Dies ist beim MPA das Innenprodukt

$$C(s(t), g(t)) = |\langle s(t), g(t)\rangle| \,. \tag{1}$$

Daraus berechnet der MPA M Koeffizienten λ_m zur linearen Signalrekonstruktion des Signals $s(t)$ mit

$$s(t) = \sum_{m=1}^{M} \lambda_m g_m(t) + R_M(t) \tag{2}$$

und $R_M(t)$ als Residuum, welches die verbleibenden Signalanteile enthält [7]. Bei unendlich langer Zerlegung konvergiert die Energie des Residuums gegen null. Somit kann die Approximation des Signals beliebig genau werden.

Die Dekomposition erfolgt iterativ, wobei in jeder Iteration m das aktuell ähnlichste Atom $\hat{g}_m(t)$ aus der Menge aller Atome $g(t) \in \mathcal{D}$ durch

$$\hat{g}_m(t) = \arg\max_{g(t)\in\mathcal{D}} |\langle R_{m-1}(t), g(t)\rangle| \tag{3}$$

berechnet und aus dem Signal extrahiert wird. Der Faktor λ_m entspricht dann genau dem Innenprodukt

$$\lambda_m = \langle R_{m-1}(t), \hat{g}_m(t)\rangle \,. \tag{4}$$

Somit berechnet sich das aktuelle Residuum durch

$$R_m(t) = R_{m-1}(t) - \lambda_m \hat{g}_m(t) \,. \tag{5}$$

Zur Erfüllung der Energieerhaltung

$$\|s(t)\|^2 = \sum_{m=1}^{M} \lambda_m^2 + \|R_M(t)\|^2 \tag{6}$$

muss die Energie jedes Atoms auf eins normiert sein.

Das Abbruchkriterium des MPA wird durch das Signal-zu-Residuum-Verhältnis (SRV) gegeben, welches durch den Quotienten der Energien des Ursprungssignals $s(t)$ zum Residuum $R_m(t)$ der Iteration m gebildet wird:

$$SRV(m) = \frac{\|s(t)\|^2}{\|R_m(t)\|^2} \,. \tag{7}$$

Ist das SRV größer als ein vorgegebener Wert δ_{SRV}, wird die Zerlegung in Atome abgebrochen und M gleich der letzten Iteration gesetzt.

2.2 Atome

Der allgemeine MPA verwendet Gabor-Atome zur Signalapproximation. Gabor-Atome $g_{s,u,f}(t)$ sind durch

$$g_{s,u,f}(t) := \frac{1}{\sqrt{s}}\, w\left(\frac{t-u}{s}\right) \mathrm{e}^{\mathrm{j}2\pi f(t-u)} \tag{8}$$

gegeben [7], wobei f die Frequenz einer komplexen Schwingung, s die Skalierung und u die zeitliche Verschiebung des Gabor-Atoms ist [6]. Die komplexe Schwingung wird mit einem Fenster $w(t)$ multipliziert.

Für Musiksignale eignen sich aufgrund der Charakteristik von Musiktönen mit Grund- und Oberschwingungen dagegen harmonische Atome. Diese sind für die Zerlegung von Signalen konzipiert, bei denen die Schwingungen einer Grundfrequenz f_0 gemeinsam mit Oberschwingungen (Partialen) der Frequenzen $f_k \approx k f_0$ betrachtet werden. Ein harmonisches Atom $h(t)$ kann als Überlagerung von K Gabor-Atomen g_{s,u,f_k} aufgefasst werden, wodurch sich

$$h(t) := \sum_{k=1}^{K} c_k \, g_{s,u,f_k}(t) \tag{9}$$

mit der Bedingung

$$\|h(t)\|^2 = \int |h(t)|^2 \, \mathrm{d}t = 1 \tag{10}$$

ergibt. Alle berücksichtigten Gabor-Atome spannen den harmonischen Unterraum auf, der durch

$$\mathcal{V}_{s,u,f_1,,,,,f_K} := \mathrm{span}\{g_{s,u,f_k}(t), 1 \le k \le K\} \tag{11}$$

beschrieben ist [6]. Um die Bedingung des MPA aus Gleichung (10) zu erfüllen, müssen die Gabor-Atome im harmonischen Unterraum quasiorthogonal zueinander sein, was mit der Näherung $f_k \approx k f_0$ und der Einhaltung einer unteren minimalen Frequenz $f_{\min}$ gegeben ist [6][2].

2.3 Umsetzung eines schnellen MPA

Ein harmonisches Wörterbuch $\mathcal{D}_\mathrm{h}$ mit vielen harmonischen Atomen h nach Gleichung (9) hat einen großen Umfang, da jedes Atom mit verschiedenen Werten s und u skaliert bzw. zeitverschoben wird. Wörterbücher mit variierenden Fensterdesigns umfassen deshalb mehrere Milliarden Atome. Dadurch und aufgrund der vielen Korrelationen im MPA benötigt die definitionsgemäße Implementierung sehr hohe Speicher- und Rechenkapazitäten. Zur Lösung dieses Problems wurde der MPA mit dem *Matching Pursuit Toolkit* (MPTK) umgesetzt [9].

Die Vorgehensweise deckt sich mit der Definition des MPA, aber die Zeitverschiebung u der Atome wird durch eine Fensterverschiebung im betrachteten Signal durchgeführt. Somit muss das Innenprodukt nur Abschnitt für Abschnitt mit den Atomen aus dem Wörterbuch durchgeführt werden, wobei jeder Abschnitt die durch die Skalierung s angegebene Länge besitzt. In diesem Fall werden die Innenprodukte über den Frequenzbereich per diskreter Fourier-Transformation (DFT) berechnet. Anschließend werden sie zwischengespeichert und die Parameter des Atoms mit dem größten Innenprodukt bestimmt. Die detaillierten Berechnungsschritte dazu sind in [6] geschildert.

In der nächsten Iteration muss das Innenprodukt nur an der Stelle aktualisiert werden, an der das im vorherigen Schritt extrahierte Atom lokalisiert war. Für die anderen Signalteile können die zwischengespeicherten Innenprodukte übernommen werden, wodurch die benötigte Rechenleistung stark verringert wird.

3 Umsetzung des HMPA zur Notenseparation

Insgesamt ist das vorgestellte Verfahren in zwei Schritte unterteilt. Zunächst werden alle im Musiksignal vorkommenden Notenspektren geschätzt und daraus die für den HMPA notwendige Bibliothek aufgebaut. Im zweiten Schritt wird das Musiksignal mithilfe der harmonischen Atome der Bibliothek durch den HMPA approximiert und die gefundenen Noten separiert.

3.1 Vorschätzung der Notenspektren

Zur Eingrenzung des Umfangs an harmonischen Atomen im Wörterbuch und als Vorgabe für den HMPA werden die Notenspektren, im Folgenden auch als Partial-Einhüllende bezeichnet, mit einem Glätteansatz und der Korrelation mit Amplitudenmodellen vorgeschätzt. Die Umsetzung der Vorschätzung kann aus dem Flussdiagramm in Abbildung 1 entnommen werden. Zuerst wird eine Kurzzeit-Fourier-

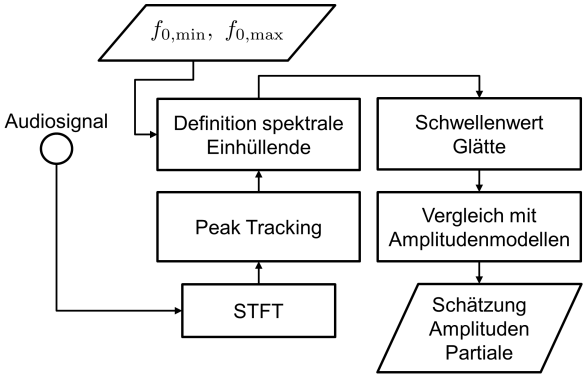

Abb. 1: Flussdiagramm zur Vorschätzung der Partiale.

Transformation (STFT) des Signals durchgeführt. Die dabei verwendeten Parameter wurden angelehnt an [2] auf

ein Hann-Fenster der Länge 128 ms bei einer Verschiebung von 10 ms festgelegt. Aus dem Zeit-Frequenz-Spektrum der STFT werden mithilfe der Methode aus [12] die Spitzenfrequenzen pro STFT-Abschnitt ermittelt.

Über den festgelegten Grundfrequenzbereich $[f_{0,\min}, f_{0,\max}]$ wird für jede mögliche Note ein Raster gebildet, vergleichbar mit einem Kammfilter, um die Amplituden der kontinuierlichen Schwingungen den Vielfachen k der Notengrundfrequenz f_0 zuzuordnen. Dabei werden die Rasterfelder durch das Intervall $[2^{-1/24}kf_0, 2^{1/24}kf_0]$ abgesteckt, um die Toleranzgrenze genau zwischen zwei Noten zu setzen. Damit werden auch Modulationen im Signal, wie z. B. das Vibrato bei einer Violine, berücksichtigt, sofern diese Amplitudenabweichung die Toleranzgrenze nicht übersteigt. Durch das Anwenden des Rasters werden die spektralen Einhüllenden definiert.

Mit dem Glättemaß γ_n aus [2] werden die glattesten Partialverteilungen aus N_i Schätzungen für eine Note i gewählt. Um mehrere Spielvariationen oder Instrumente pro Note in Betracht zu ziehen, wird in dieser Arbeit nicht ausschließlich der beste Kandidat verwendet, sondern eine Auswahl der glattesten Kandidaten $\mathcal{E}_\mathrm{g}$ ermittelt, deren Glätte über dem Schwellenwert δ_g liegt.

Ausgehend von typischen Modellgruppen $\mathcal{M}$ für Instrumente nach [17] können fundamentale Modelle $\boldsymbol{e}_\mathrm{ref} \in \mathcal{M}$ als Referenz für die Wahl korrekter Partialamplituden genutzt werden. Um die zuvor über die Glätte gewählten Partialamplituden $\mathcal{E}_\mathrm{g}$ mit den Referenzen aus den fundamentalen Modellen zu vergleichen, wird der Korrelationskoeffizient ρ nach Pearson [1] berechnet. Es werden die Partialamplituden zur Schätzung verwendet, die eine ausreichende Ähnlichkeit zu den Referenzmodellen aufzeigen. Ein Beispiel für die resultierende Schätzung aller möglichen Partialverteilungen in einem Signal ist in Abbildung 2 über ein Relief dargestellt.

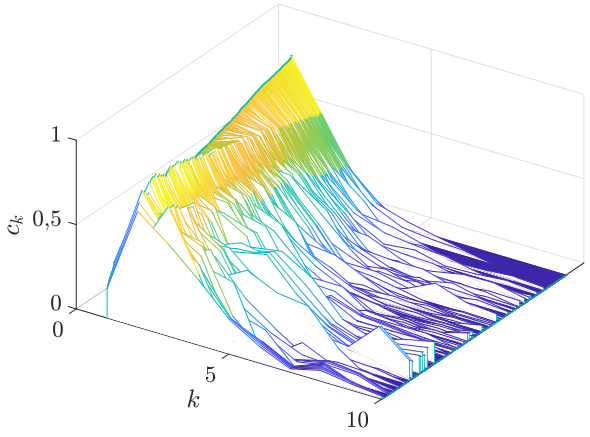

Abb. 2: Partialschätzung zu J. Brahms – Horn Trio in Es-Dur, Op. 40; normiert auf $\sum_k |c_k|^2 = 1$.

3.2 Konstruktion des Wörterbuchs

Die Konstruktion des Wörterbuchs setzt sich aus den Amplitudenschätzungen der Partiale, variierenden Atomlängen und den Fensterfunktionen zusammen. Trotz der Schätzung der Partiale aus dem Musiksignal sind dabei die Voraussetzungen für einen MPA, wie die Forderung nach Überdefiniertheit des Wörterbuchs und die Quasiorthogonalität des harmonischen Unterraums, gegeben.

3.2.1 Partialschätzungen

Anhand der Informationen über die Amplitudenverteilung der geschätzten Elemente $\boldsymbol{e}_\mathrm{est} \in \mathcal{E}_\mathrm{est}$ für die Partiale wird das Wörterbuch mit harmonischen Atomen gefüllt. Aus Abschnitt 2.2 ergibt sich die Form

$$h(t) = \sum_{k=1}^{K} c_k \frac{1}{\sqrt{s}}\, w\left(\frac{t-u}{s}\right) \mathrm{e}^{\mathrm{j}2\pi f_k (t-u) + \mathrm{j}\phi_k} \tag{12}$$

mit der Erweiterung um die Phasen der k-ten Partiale $\phi_k \in [0, 2\pi)$. Die Errechnung der Phasenlage der Partiale ist mit der schnellen Umsetzung des HMPA gegeben, die in Abschnitt 2.3 beschrieben wird. Um die Linearfaktoren c_k des harmonischen Unterraums $\mathcal{V}_{s,u,f_1,\ldots,f_K}$ an die geschätzten Partiale $\boldsymbol{e}$ anzupassen, müssen die Partiale auf die Energie von 1 normiert werden. Daraus ergeben sich die Faktoren

$$c_k = \frac{e_k}{\sqrt{\|\boldsymbol{e}\|^2}} \tag{13}$$

und die Energieerhaltung aus Gleichung (6) ist erfüllt.

Die Skalierung s gibt die Länge der Atome an und wird in der Umsetzung durch die Fensterlänge bestimmt. Bei einer Schrittweite von 31,3 ms sind die verwendeten variablen Fenster 31,3 ms bis 250 ms lang, analog zu [2].

3.2.2 Fensterfunktionen

Angelehnt an die Ergebnisse aus den Arbeiten [6] und [2] werden spezielle Fensterfunktionen verwendet, um gleichermaßen symmetrische und asymmetrische Muster im Signal adäquat extrahieren zu können. Die verschiedenen Muster entstehen durch unterschiedliche Instrumentencharakteristiken. In dieser Arbeit werden das Gauß-Fenster zur klassischen Gabor-Zerlegung sowie Fensterdesigns nach Hamming, Hann, Cosinus und Blackman als symmetrische Fenster im Wörterbuch verwendet.

Zur Nachbildung von Anschlagsphasen oder langsam ausklingenden Tönen wird das Fof-Fenster als asymmetrisches Fenster mit den Parametern aus [6] eingesetzt.

3.3 Nachverarbeitung

Bei der Nachbereitung wird zwischen Transkription und Separation unterschieden. Die Transkription ist ein klar quantisierter Vorgang mit Definition von Grundfrequenzen sowie Anfangs- und Endzeitpunkten der gespielten Noten. Im Gegensatz dazu ist die Separation von Melodien eine Zuordnung von Signalanteilen zu den auftretenden Noten. Gerade in der Darstellungsform über harmonische Atome wird in der Regel nicht von einer reinen Atom-Note-Äquivalenz ausgegangen, sondern mehrere Atome bilden die zeitliche Charakteristik der Note ab [10].

3.3.1 Transkription

Zur Transkription werden iterativ alle quantisierten Atome zu einem Cluster zusammengefasst, deren Überschneidungsrate größer als $\delta_{t,OR} = 15\,\%$ ist. Die Überschneidungsrate *OR* zweier Atome mit den Startzeitpunkten $t_{\mathrm{on},1}$, $t_{\mathrm{on},2}$ und Endzeitpunkten $t_{\mathrm{off},1}$, $t_{\mathrm{off},2}$ berechnet sich dabei nach [14] durch

$$OR_{1,2} = \frac{\min(t_{\mathrm{off},1}, t_{\mathrm{off},2}) - \max(t_{\mathrm{on},1}, t_{\mathrm{on},2})}{\max(t_{\mathrm{off},1}, t_{\mathrm{off},2}) - \min(t_{\mathrm{on},1}, t_{\mathrm{on},2})}\,. \qquad (14)$$

Die erhaltenen Atomcluster werden anschließend anhand der Zeitdauer gefiltert. Die untere Grenze für eine gespielte Note ist auf $\Delta_t = 80\,\mathrm{ms}$ festgelegt. Das entspricht in etwa der Hälfte der kürzesten Noten in den Datensätzen.

3.3.2 Separation

Durch die quantisierte Grundfrequenz f_0 der extrahierten HMPA-Atome zu Noten westlicher Musik können die Atome den Noten einer Melodie zugeordnet werden.

In dieser Arbeit wird davon ausgegangen, dass die Referenz der Noten einer Melodie vorhanden ist. Ausgehend von dieser Referenz wird überprüft, welche Atome durch eine Note dieser Melodie eingeschlossen werden. Da die annotierten Noten der Melodien zeitlich von *Onset* t_{on} und *Offset* t_{off} der real gespielten Noten abweichen können, wird eine Toleranz von $\delta_{t,\mathrm{sep}} = 100\,\mathrm{ms}$ verwendet. Der Wert für die Toleranz ist relativ hoch, da die Instrumente häufig noch nachklingen oder leicht früher als die Referenz gespielt werden. Mit einer durchschnittlichen minimalen Tonlänge von 130 ms des TRIOS-Datensatzes aus Abschnitt 4.1 liegt die Toleranz aber noch unter der Länge einer gespielten Note und kann so etwaige kurz davor gespielten Töne nicht fälschlicherweise einschließen.

4 Ergebnisse

Das in dieser Arbeit beschriebene Verfahren wird anhand von einem Piano-Datensatz und einem Trio-Datensatz unterschiedlicher Instrumente mithilfe der in Abschnitt 4.2 erläuterten Gütemaße validiert.

Nach jeder Iteration des HMPA wird das Abbruchkriterium SRV nach Gleichung (7) überprüft. In empirischen Versuchen stellte sich heraus, dass ein Schwellenwert δ_{SRV} von 8 dB für Stücke mit einem Instrument und 12 dB für Stücke mit drei Instrumenten gut geeignet ist.

4.1 Datensätze

Zur Untersuchung der Transkriptionsgüte mit dem HMPA wird zum einen, angelehnt an [2], ein Datensatz mit einem synthetisierten Piano als einziges Instrument verwendet. Um die Transkription für die Schätzung mehrerer Instrumente unter realen Bedingungen zu evaluieren, wurde der TRIOS-Datensatz [5] mit Stücken, in denen drei Instrumente spielen, in die Auswertung aufgenommen. Anhand von diesem Datensatz wird neben der Transkription auch die Separation evaluiert.

Die ausgewählten Stücke aus dem TRIOS-Datensatz beinhalten keine perkussiven Elemente und das Piano ist in jedem Stück das Grundinstrument. Zusätzlich zum Piano spielen Streichinstrumente sowie Holz- und Blechblasinstrumente. Es werden folgende Kombinationen betrachtet:

- Klarinette, Bratsche und Piano (Mozart),
- Horn, Violine und Piano (Brahms),
- Fagott, Trompete und Piano (Lussier),
- Cello, Violine und Piano (Schubert).

4.2 Gütemaße

Zur Untersuchung der Transkriptionsgüte wird die Evaluationsmethode der MIREX-Wettbewerbe [3] verwendet, welche in [4] und [14] beschrieben ist. Dabei wird die Anzahl der korrekt klassifizierten Noten mit N_{corr}, die gesamte Anzahl erkannter Noten mit N_{est} und die Gesamtzahl an *Ground-Truth*-Noten mit N_{ref} gegeben. Alle Bedingungen für eine korrekt klassifizierte Note wurden mithilfe von [4] nach der MIREX-Konvention umgesetzt. Die Gütemaße sind *Precision* P, *Recall* R und das kombinierte *F-Measure* F_1 mit

$$P = \frac{N_{\mathrm{corr}}}{N_{\mathrm{est}}}\,, \quad R = \frac{N_{\mathrm{corr}}}{N_{\mathrm{ref}}} \quad \text{und} \quad F_1 = 2 \cdot \frac{P \cdot R}{P + R} \qquad (15)$$

Diese Arbeit führt darüber hinaus eine Separation der einzelnen, von verschiedenen Instrumenten gespielten Me-

lodien durch. Die verwendeten Stücke setzen sich aus je drei gemischten Instrumentenspuren zusammen. Die einzelnen Spuren sind mit einer Melodie-Note-Validierung versehen und werden mithilfe der Umsetzung aus Abschnitt 3.3.2 den harmonischen Atomen zugeordnet. Zur Evaluation der Separation wird, wie in der Signaltheorie üblich, wenn die Ähnlichkeit zwischen zwei Signalen gemessen wird, Pearsons Korrelationskoeffizient ρ verwendet [1]. Der Korrelationskoeffizient ρ zwischen dem rekonstruierten Signalvektor $\mathbf{x}_{\text{rec}}$ und dem Referenzsignal $\mathbf{x}_{\text{ref}}$ (einzeln gespielte Spur im Separationsdatensatz TRIOS) mit jeweils der Länge L berechnet sich zu

$$\rho(\mathbf{x}_{\text{rec}}, \mathbf{x}_{\text{ref}}) = \frac{\text{cov}(\mathbf{x}_{\text{rec}}, \mathbf{x}_{\text{ref}})}{\sigma_{\mathbf{x}_{\text{rec}}} \sigma_{\mathbf{x}_{\text{ref}}}} . \tag{16}$$

4.3 Auswertung

Zunächst werden die Ergebnisse der Transkription diskutiert. Anschließend folgt die Validierung der Separation.

4.3.1 Ergebnisse der Transkription

Die am *F-Measure*-Wert gemessene Qualität der Transkription liegt für den Piano-Datensatz bei durchschnittlich 0,45 mit einer Standardabweichung von 0,17. Daraus folgt, dass die Genauigkeit zwischen den Stücken stark variiert. Das kann unter anderem darauf zurückgeführt werden, dass Atome nur geclustert werden können, wenn sie sich zeitlich ausreichend überschneiden. Dies ist vor allem bei langen Noten häufig nicht der Fall. Zur Verdeutlichung wurde die durchschnittliche Noten-Spieldauer der Stücke im Pianodatensatz errechnet. Anhand der berechneten Notenlängen wurde der Datensatz in zwei Gruppen unterteilt. Die Verteilung der F_1-Werte der Gruppen ist durch den Boxplot in Abbildung 3 dargestellt.

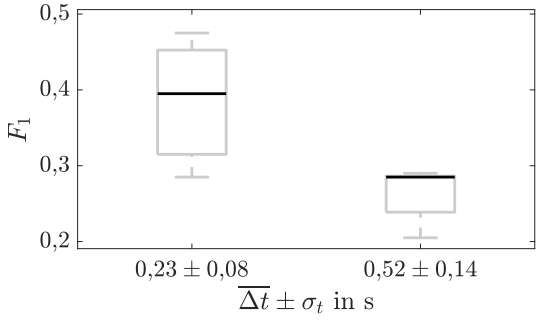

Abb. 3: *F-Measures* F_1 der Piano-Transkription zum Durchschnitt und der Standardabweichung der Notenspieldauer.

Unter Berücksichtigung des Einflusses der Notenlänge auf den F_1-Wert zeigt der HMPA keine signifikant schlechtere Transkription bei drei Instrumenten als bei einem Instrument. Dieser Zusammenhang wird anschaulich im Diagramm in Abbildung 4 dargestellt. Dazu wird der

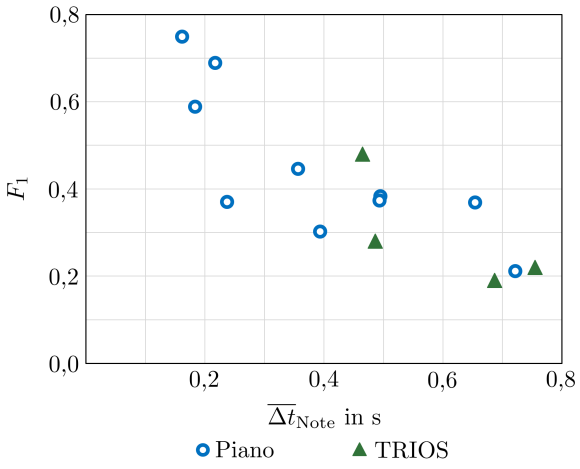

Abb. 4: *F-Measure* F_1 über die durchschnittliche Notenlänge $\overline{\Delta t}_{\text{Note}}$ der Stücke aus dem Piano- und dem TRIOS-Datensatz.

F_1-Wert über die durchschnittlichen Notenlängen $\overline{\Delta t}_{\text{Note}}$ der Piano- und TRIOS-Stücke aufgezeichnet. Die TRIOS-Stücke besitzen im Durchschnitt einen höheren Transkriptionsfehler, allerdings werden in den TRIOS-Stücken auch durchschnittlich längere Töne gespielt. Verglichen mit den Punkten der Pianostücke liegen die der TRIOS-Stücke in Abbildung 4 in vergleichbaren Bereichen.

4.3.2 Ergebnisse der Separation

Zur Auswertung der Separation polyphoner Melodien wird der TRIOS-Datensatz mit drei parallelen Melodien verwendet. Die Separation wird durch den HMPA am gemischten Gesamtsignal durchgeführt. Mit dem Wissen über die einzelnen Melodieverläufe werden die HMPA-Atome zugeordnet und das Signal separiert. Anschließend werden die separierten Melodien anhand der Korrelationskoeffizienten ρ nach Gleichung (16) mit den separaten Originalspuren verglichen. Die Ergebnisse sind in Tabelle 1 aufgeführt.

Der durchschnittliche Korrelationskoeffizient liegt bei $\overline{\rho} = 0{,}608$. Es gibt starke Unterschiede bei der Separation der einzelnen Melodien, die von jeweils einem anderen Instrument gespielt werden. So ist die Extraktion der Klarinette beim Stück von Mozart mit einem Korrelationskoeffizient von $\rho = 0{,}923$ das am besten rekonstruierte Signal. Den zweitbesten Wert stellt die Rekonstruktion

Tab. 1: Korrelationskoeffizient ρ der separierten Melodiespuren mit dem jeweiligen Original.

Komponist – Stück	Instrument		
Wolfgang A. Mozart	**Klarinette**	**Bratsche**	**Piano**
K. 498	0,923	0,525	0,534
Johannes Brahms	**Horn**	**Violine**	**Piano**
Op. 40	0,688	0,440	0,570
Mathieu Lussier	**Fagott**	**Trompete**	**Piano**
Op. 8	0,236	0,670	0,513
Franz Schubert	**Cello**	**Violine**	**Piano**
D. 929. Op. 100	0,780	0,594	0,818

des Pianos beim Stück von Schubert mit $\rho = 0{,}818$ dar. Die schlechtesten Rekonstruktionen werden bei der Violine mit $\rho = 0{,}44$ aus dem Stück von Brahms sowie dem Fagott beim Stück von Lussier mit $\rho = 0{,}236$ erreicht. Ursachen der Korrelationsunterschiede bei den separierten Melodien sind zum einen unterschiedliche Anteile der Melodien an der Gesamtenergie des Musiksignals sowie Einflüsse der Instrumentencharakteristik.

5 Zusammenfassung

Zur Separation von Noten eines polyphonen monauralen Musiksignals wird ein harmonischer Matching-Pursuit-Algorithmus eingesetzt, dessen Bibliothek vor der Separation durch Amplitudenschätzungen der Tonspektren gebildet wird. Im Vergleich zur Transkription von Musiksignalen eines Instruments wird bei mehreren gleichzeitig gespielten Instrumenten eine ähnlich gute Transkription erreicht.

Mit Vorwissen über die Melodieverläufe einzelner Stimmen lassen sich die extrahierten Notenelemente zu separierten Melodien dieser Stimmen zusammenfassen. Diese rekonstruierten Einzelsignale wurden durch Korrelationskoeffizienten validiert, wobei gute Übereinstimmungen mit dem Originalsignal festgestellt wurden.

In zukünftigen Arbeiten soll die Amplitudenschätzung der Partiale mithilfe von Methoden des *Machine Learnings* untersucht werden, da diese die trainierbaren Charakteristika von Notenmodellen sehr gut abbilden können. Darüber hinaus kann der Einfluss der Extraktionsreihenfolge approximierter Noten analysiert werden.

Literatur

[1] J. Benesty, J. Chen, Y. Huang und I. Cohen. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*, S. 1–4. Springer, 2009.

[2] J. J. Carabias-Orti, P. Vera-Candeas, F. J. Cañadas-Quesada und N. Ruiz-Reyes. Music scene-adaptive harmonic dictionary for unsupervised note-event detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):473–486, 2010.

[3] J. Downie, K. West, A. Ehmann und E. Vincent. The 2005 music information retrieval evaluation exchange (mirex 2005): Preliminary overview. In *ISMIR*, S. 320–323, 2005.

[4] Z. Duan, B. Pardo und C. Zhang. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2121–2133, 2010.

[5] J. Fritsch. High quality musical audio source separation. Masterthesis, Centre for Digital Music, 2012.

[6] R. Gribonval und E. Bacry. Harmonic decomposition of audio signals with matching pursuit. *IEEE Transactions on Signal Processing*, 51(1):101–111, 2003.

[7] R. Gribonval, P. Depalle, X. Rodet, E. Bacry und S. Mallat. Sound signals decomposition using a high resolution matching pursuit. In *Proceedings of the International Computer Music Conference*, S. 293–296, 1996.

[8] P.-S. Huang, M. Kim, M. Hasegawa-Johnson und P. Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23 (12):2136–2147, 2015.

[9] S. Krstulovic und R. Gribonval. MPTK: Matching pursuit made tractable. In *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Band 3, S. III496–III499, 2006.

[10] P. Leveau, E. Vincent, G. Richard und L. Daudet. Instrument-specific harmonic atoms for mid-level music representation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):116–128, 2008.

[11] S. G. Mallat und Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[12] R. McAulay und T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, 1986.

[13] A. Rizzi, M. Antonelli und M. Luzi. Instrument learning and sparse NMD for automatic polyphonic music transcription. *IEEE Transactions on Multimedia*, 19(7):1405–1415, 2017.

[14] M. P. Ryynanen und A. Klapuri. Polyphonic music transcription using note event modeling. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, S. 319–322. IEEE, 2005.

[15] S. Sigtia, E. Benetos und S. Dixon. An end-to-end neural network for polyphonic piano music transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):927–939, 2016.

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

[17] C. Yeh und A. Roebel. The expected amplitude of overlapping partials of harmonic sounds. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, S. 3169–3172. IEEE, 2009.