Exemplarische Untersuchung eines Quellseparationsalgorithmus

Bachelorarbeit aus Aufnahmetechnik 1, SE

Leon Kaiser

Betreuung: Univ.Prof. DI Dr. Alois Sontacchi

Graz, 5. Oktober 2020

institut für elektronische musik und akustik

Zusammenfassung

In dieser Arbeit wird exemplarisch ein Quellseparationsalgorithmus untersucht, um einen Einstieg in die Thematik Quellseparation zu erleichtern. Dafür werden erst die theoretischen Grundlagen erläutert. Es wird dabei auf das Quelle-Filter-Modell, die Zeit-Frequenz-Transformation, die Matrixfaktorisierung, eine cepstrale Repräsentation und auf k-means eingegangen. Daraufhin beschreibt die Arbeit die prozedualen Schritte des Algorithmus.
Es folgt eine Analyse von den Ergebnissen, die der Algorithmus liefert. Dafür wurden die Signalverarbeitungskonzepte von [EVHH11] und [Vin12] herangezogen, welche für die Analyse von Quellseparationen entwickelt wurden. Im Genaueren werden Ergebnisschwankungen und Parametrierbarkeit untersucht. Es konnte keine Parametrierbarkeit festgestellt werden. Es wurde herausgefunden, dass der Einfluss von zwei untersuchten Parametern gering ist.

Inhaltsverzeichnis

1 Einleitung **4**

2 Übersicht über den Algorithmus **5**

2.1 Theoretische Grundlagen 5

2.1.1 Quelle-Filter-Modell 5

2.1.2 Kurz-Zeit-Fourier-Transformation (STFT) 5

2.1.3 Nicht-negative Matrix-Faktorisierung (NMF) 7

2.1.4 Mel-Frequenz-Cepstrum-Koeffizienten (MFCC) 9

2.1.5 K-Means 11

2.2 Vorgehensweiße und prozeduale Schritte 12

3 Evaluierung der Parametrierbarkeit **14**

3.1 Ziele der Evaluierung 14

3.2 Evaluierungsprozess 15

3.3 Evaluierung der Ergebnisschwankungen 16

3.4 Evaluierung von Voreinstellungen 18

4 Diskussion und Schlussfolgerungen **19**

1 Einleitung

Um Quellseparation zu verstehen, muss definiert werden was man als Quellen versteht. Diese Arbeit beschränkt sich auf Akustik Signale. Quellen bezeichnen in diesem Kontext Schallquellen (z.B. Instrumente oder Sprecher). Damit ein Algorithmus mit Schallquellen arbeiten kann, müssen diese digitalisiert werden. So erhält man die digitalen Quell-Signale $s_m[n]$.

$$x[n] = \sum_{m=1}^{M} s_m[n] \tag{1}$$

Da nun definiert ist was Quellen sind, kann erklärt werden was Quellseparation bedeutet. Man versteht unter Quellseparation aus einer Signalmischung $x[n]$ (Gleichung 1) die einzelnen Quellsignale $s_m[n]$ heraus zu trennen. Man unterscheidet zwischen einem Vorgehen, für welches Zusatzinformation über die Quellen miteinbezogen wird und einem Vorgehen ohne Informationen über die Quellen. Einen Ansatz bei dem nur die ursprüngliche Signalmischung zur Verfügung steht, nennt man einen blinden Ansatz ('blind-source-separation' - BSS). Die hier betrachtete Methode ist eine BSS, mit der Ausnahme, dass die Anzahl der Quellen bekannt sein muss. Die Separation muss deshalb, da keine anderen Anhaltspunkte gegeben sind, mithilfe von Signalmerkmalen von $x[n]$ durchgeführt werden.
Menschen können eine Quellseparation aufgrund von Signalmerkmalen durchführen. Dies passiert zum Beispiel, wenn man zwei Musikinstrumente gleichzeitig hört. Man kann die Instrumente unterscheiden und sich auf eines fokussieren. Es gibt unterschiedliche Signalmerkmale die dafür hilfreich sind. Zum Beispiel können Pegel- und Phasendifferenzen zwischen rechtem und linkem Ohr helfen. Ein weiteres hilfreiches Signalmerkmal sind die unterschiedlichen, charakteristischen Klangfarben von verschiedenen Quellen. Durch Klangcharakteristiken kann man auch Quellen unterscheiden wenn sie von der selben Position abstrahlen. So ist das der Fall beim Abspielen von monauralen Aufnahmen.
Ein Ansatz für die Realisation von Quellseparationsalgorithmen ist es, diese Signalmerkmale heranzuziehen. In [ME07] wird ein Algorithmus beschrieben, der auf interauralen Zeit- und Phasendifferenzen basiert. Der Algorithmus, der in dieser Arbeit untersucht wird, arbeitet mit Klangfarben [SG09].
Nachdem geklärt ist, was Quellseparation bedeutet, stellt sich die Frage, wofür sie gebraucht wird. Quellseparation kann für Spracherkennungsmethoden [PJLL99], Instrumentenklassifizierung [HKV09] oder auch zur Weiterentwicklung von Cochlea-Implantate [KL08] hilfreich sein. Außerdem ist es denkbar, bei guter Separation die neue Zusammenmischung mit verbesserter Balance und Durchhörbarkeit der einzelnen Instrumente bei monauralen Aufnahmen zu ermöglichen.

2 Übersicht über den Algorithmus

2.1 Theoretische Grundlagen

Um den Algorithmus verstehen zu können, wird vor der Erläuterung der prozedualen Schritte (siehe Sektion 2.2) auf für das Verständnis wichtiger Thematiken eingegangen.

2.1.1 Quelle-Filter-Modell

Im Quelle-Filter Modell wird ein Signal $s[n]$ als gefiltertes Quellsignal betrachtet. Im Kontext dieses Modells hat das Wort Quellsignal eine andere Bedeutung als im Kontext der Quellseparation. Um Verwirrung zu vermeiden wird das Quellsignal des Quelle-Filter-Modells im folgenden als Anregungssignal bezeichnet.
Das Modell wurde ursprünglich zur Sprachanalyse entworfen [Fan70]. Dabei wird die Glottis als Erzeuger des Anregungssignals $e[n]$ betrachtet und der Vokaltrakt als Filter welcher das Anregungssignal formt. Die Impulsantwort des Filters wird hier als $h[n]$ definiert.

$$s[n] = e[n] * h[n] \tag{2}$$

Mittlerweile wird das Modell auch auf Musik angewendet. So verwenden zum Beispiel [HKV09] und [NSHAG07] ein Quelle-Filter Modell zur Instrumentenklassifizierung.
In dieser Arbeit wird das Modell leicht modifiziert angesetzt. Dabei müssen Anregungssignal und Filter nicht physikalisch getrennt vorhanden sein. Das Modell dient hier einer abstrakten Signalanalyse. Das Anregungssignal $e[n]$ beinhaltet die Information welche Frequenzen im Instrumentalsignal $s[n]$ auftreten. Alle diese Frequenzen treten in $e[n]$ mit der gleicher Amplitude auf. Der Filter $h[n]$ stellt somit im Frequenzbereich als die Hüllkurve des Spektrums von $s[n]$ dar. Der Filter $h[n]$ gewichtet also die Frequenzen von $e[n]$ und beinhaltet somit die Information über die Klangcharakteristik (siehe Abbildung 1).

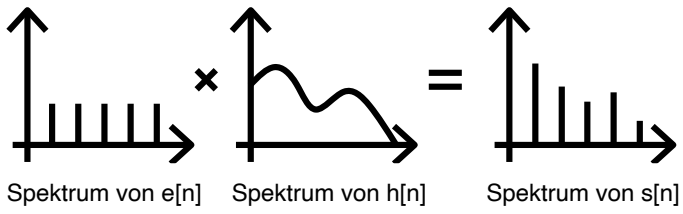

Abbildung 1 – Quelle-Filter-Modell für abstrakte Signalanalyse (vgl. [NSHAG07] Fig.1)

2.1.2 Kurz-Zeit-Fourier-Transformation (STFT)

Eine STFT transformiert ein Signal in die Zeit-Frequenz-Domäne. Man kann somit betrachten, wie sich das Spektrum des Signals mit der Zeit wandelt.

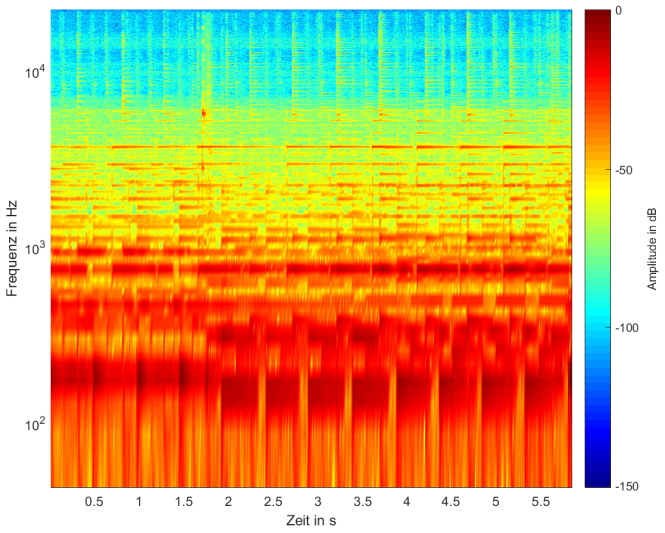

Abbildung 2 – Darstellung einer STFT-Matrix $\mathbf{X}$ (Spektrogramm)

Im Folgenden wird von einem Signal $x[n]$, einem Fenstersignal $w[n]$ und einem resultierenden Signal $X[n,k]$ ausgegangen.

$$x[n] \in \mathbb{R} \; \forall \, n \in [0, N_x - 1] \tag{3}$$
$$w[n] \in \mathbb{R} \; \forall \, n \in [0, N_w - 1] \tag{4}$$
$$X[n,k] \in \mathbb{C} \; \forall \, n \in [0, T-1], k \in [0, K-1] \tag{5}$$

Es werden mit dem Fenster $w[n]$ zu T Zeitpunkten Signalausschnitte von $x[n]$ gebildet. Für diese Signalausschnitte wird eine diskrete Fourriertransformation (DFT) durchgeführt. Die STFT (Gleichung 7) besteht also aus T DFTs (Gleichung 6).

$$X[k] = \sum_{n=0}^{N_x - 1} x[n] e^{-\frac{j2\pi}{N_x} kn} \tag{6}$$

In Gleichung 7 ist ein allgemeiner Ansatz einer diskreten STFT zu sehen.

$$\forall n \in [0, N_x - N_w], \forall k \in [0, N_w - 1] :$$
$$X[n,k] = \sum_{m=0}^{N_w - 1} x[n+m] w[m] e^{-\frac{j2\pi}{N_w} km} \tag{7}$$

Bei diesem allgemeinen Ansatz wird für jeden möglichen Zeitpunkt eine Fensterung und DFT durchgeführt. Die Zeitpunkte $n > N_x - N_w$ werden ausgelassen, da hier Werte außerhalb des Definitionsbereichs von $x[n]$ benötigt werden würden.
Ausgehend von dem allgemeinen Ansatz wird die STFT nun abgeändert, um überflüssige Information zu eliminieren und um Rechenleistung in der Praxissituation zu sparen.

Dadurch, dass sich für $x[n]$ auf reelwertige Signale beschränkt wird, ergibt sich eine Symmetrie im Frequenzbereich. Der Informationsgehalt bei $X[n,k]$ beschränkt sich also auf die Hälfte der Frequenzpunkte. Somit können die Frequenzpunkte für $k > \left\lceil \frac{N_w+1}{2} \right\rceil$ fallen gelassen werden. Es resultieren K Frequenzpunkte (Gleichung 8).
Weiteres wird eine "hop-size" p eingeführt um die Rechendauer zu minimieren. Anstatt für jeden möglichen Zeitpunkt eine DFT zu berechnen (Gleichung 7), wird dies für jeden p-ten Zeitpunkt durchgeführt. So ergibt sich eine geringere Anzahl an Signalausschnitten T (Gleichung 9).

$$K = \left\lceil \frac{N_w + 1}{2} \right\rceil \tag{8}$$

$$T = \left\lfloor \frac{N_x - N_w}{p} \right\rfloor \tag{9}$$

Da die Signale mit einem Algorithmus verarbeitet werden, bietet sich an $X[n,k]$ durch die Matrix $(K \times T)$-Matrix $\underline{\mathbf{X}}$ zu repräsentieren. $\mathbf{X}$ bezeichnet im Folgenden die Matrix, für die noch zusätzlich von jedem Element von $\underline{\mathbf{X}}$ der Betrag gebildet wird.

$$\mathbf{X} = \begin{bmatrix} |X[0,0]| & |X[0,1]| & ... & |X[0,T-1]| \\ |X[1,0]| & |X[1,1]| & ... & |X[1,T-1]| \\ ... & ... & ... & ... \\ |X[K-1,0]| & |X[K-1,1]| & ... & |X[K-1,T-1]| \end{bmatrix} \tag{10}$$

Zur Veranschaulichung ist in Abbildung 2 die Matrix $\mathbf{X}$ einer beliebigen Signalmischung dargestellt. Durch die Berechnung entsteht ein linearer Frequenzgang. Die Frequenzachse in Abbildung 2 ist logarithmisch. Das führt dazu, dass die Auflösung in der Abbildung bei den tiefen Frequenzen geringer ist und eine Verzerrung in der Darstellung der Frequenzpunkte zu berücksichtigen ist.

2.1.3 Nicht-negative Matrix-Faktorisierung (NMF)

NMF bezeichnet eine Gruppe von Algorithmen, deren Ziel es ist eine Matrix $\mathbf{X}$ durch ein Matrixprodukt $\mathbf{BG}$ zu approximieren, wobei die Matrizen B und G nur nicht negative Elemente haben dürfen.

$$\mathbf{X} \approx \tilde{\mathbf{X}} = \mathbf{BG} \tag{11}$$

Für dieses Ziel gibt es unterschiedliche Algorithmen. Die hier verwendete Variante und die damit verbundenen Formeln sind in [Vir07] zu finden. In diesem Kapitel wird zuerst der Ablauf grob zusammengefasst und dann darauf eingegangen, wie die Anwendung einer NMF auf digitalisierte Akustik Signale zu verstehen ist.

Ablauf des Algorithmus

Im Falle der Quellseparation, ist die STFT-Matrix $\mathbf{X}$ die zu faktorisierende Matrix. Es ist zu beachten, dass die elementweise Betragsbildung notwendig ist. Schließlich kann eine NMF nur mit einer Matrix durchgeführt werden, die reelle und nicht negative Werte enthält.

Es werden eine $(K \times I)$-Matrix $\mathbf{B}$ und eine $(I \times T)$-Matrix $\mathbf{G}$ mit Zufallswerten initialisiert. Danach beginnt der rekursive Teil des Algorithmus, wobei das Produkt $\mathbf{BG}$ an $\mathbf{X}$ angenähert wird.
Es ist zu beachten, dass aufgrund der anfänglich zufälligen Initialisierung die NMF nicht immer zum gleichen Ergebnis konvergiert. Es resultieren also bei mehrfacher Durchführung mit dem selben Signal unterschiedliche Ergebnisse.

Anwendung auf digitalisierte Akustik Signale

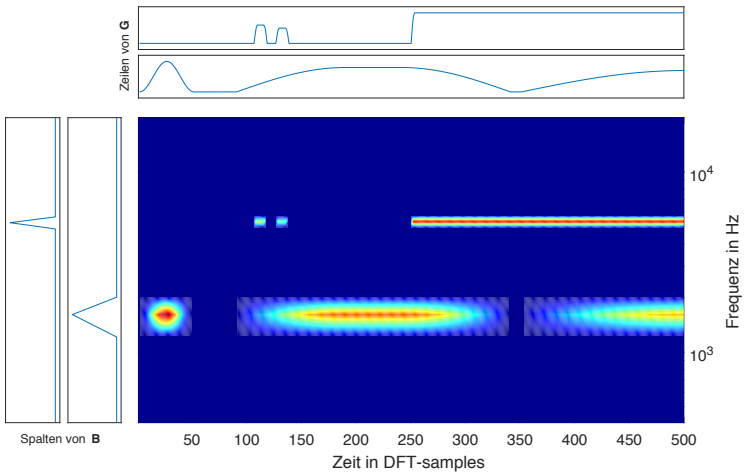

Abbildung 3 – Quelle: in Anlehnung an [Sma04] Fig. 1. - Veranschaulichung einer theoretischen NMF in der zwei Sinus-Töne getrennt werden. Hier gilt: $I = 2$.

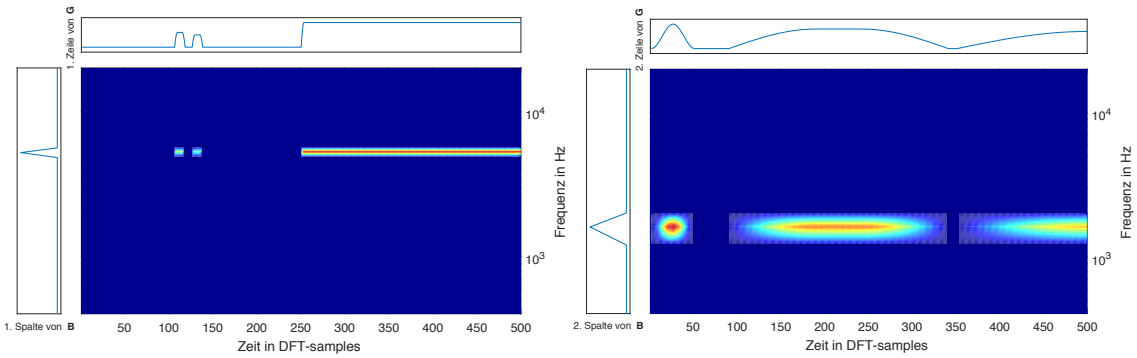

Abbildung 4 – Es werden die Komponenten $\mathbf{C}_1$ und $\mathbf{C}_2$ dargestellt. Die Komponenten wurden nach der NMF der Matrix aus Abbildung 3 gebildet.

Die STFT-Matrix $\mathbf{X}$ stellt den Amplitudenverlauf in der Zeit-Frequenzdomäne dar. Die Spalten $\mathbf{B}_i$ der Matrix $\mathbf{B}$ sind als Frequenzspektren zu verstehen. Die Spektren $\mathbf{B}_i$ werden als spektrale Komponenten des Originalsignals bezeichnet. Die spektralen Komponenten

sind Fragmente eines Spektrums, die von Zeit zu Zeit mit einer gewissen Amplitude einsetzen. Die Zeilen $\mathbf{G}_i$ der Matrix $\mathbf{G}$ sind die korrespondierenden zeitlichen Amplitudenverläufe der spektralen Komponenten (Abbildung 3).
$\mathbf{X}$ kann in einzelne Komponenten $\{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_I\}$aufgespalten werden. Die Komponenten werden gebildet indem man das Matrixprodukt zwischen der i-ten Spalte von $\mathbf{B}$ und der i-ten Zeile von $\mathbf{G}$ berechnet.

$$\mathbf{C}_i = \mathbf{B}_i \mathbf{G}_i \tag{12}$$

Es resultieren I Matrizen mit einem Rang von eins (Abbildung 4).

2.1.4 Mel-Frequenz-Cepstrum-Koeffizienten (MFCC)

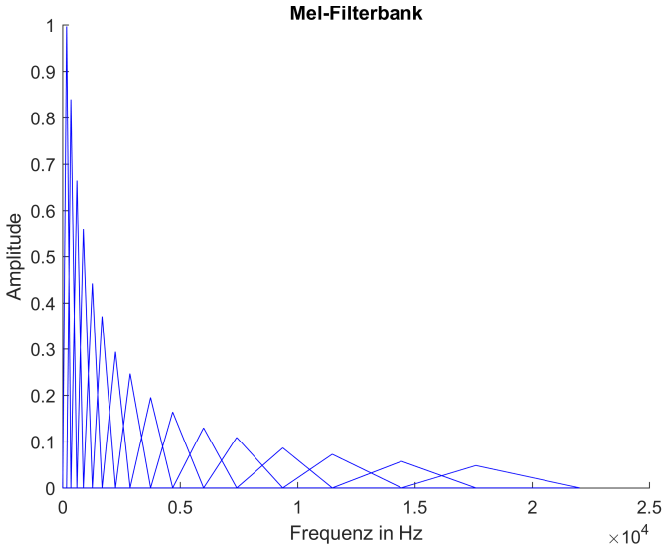

Abbildung 5 – Veranschaulichung einer Mel Filterbank $\mathbf{R}$ mit $N_{mel} = 16$

Die MFCC sind eine spezielle Repräsentation eines Spektrums. Sie werden unter anderem in [DM80] und [SG09] erläutert. Die MFCC stellen ein Spektrum des neu skalierten Frequenzspektrums dar. Ziel der Bildung der MFCC kann es sein, die Klangcharakteristik eines Signals möglichst komprimiert zu beschreiben. Die Hüllkurve eines Spektrums ist eine geeignete Art die Klangcharakteristik abzubilden (Sektion 2.1.1). MFCC können die Hüllkurve eines Spektrums in komprimierter Form darstellen.

Die MFCC werden in drei Schritten aus dem Betrag der DFT eines Signals $|X[k]|$ berechnet:

1. Anwendung einer Mel-Filterbank
2. Kompression mithilfe des Logarithmus
3. Anwendung der Diskreten Kosinus Transformation (DCT)

Anwendung einer Mel-Filterbank

Der quadrierte Betrag einer DFT $|X[k]|^2$ wird im ersten Schritt mittels einer Mel-Filterbank in Frequenzbänder zusammengefasst. Mithilfe der Filterbank werden die K Frequenzwerte der Betrags-DFT $|X[k]|$ in N_{mel} Mel-Werte $F[k_{mel}]$ zusammengefasst. Dafür wird die $N_{mel} \times K$ Matrix $\mathbf{R}$ verwendet.

$$F[k_{mel}] = \mathbf{R}|X[k]|^2 \tag{13}$$

Jede Zeile der Matrix enthält einen Dreiecksfilter der Filterbank. Es werden für die sich überlappenden Dreiecksfilter die Zentrumsfrequenzen so gewählt, dass sie in der Mel-Skala äquidistant sind (siehe Abbildung 5). Dafür wird die selbe Mel-Frequenz-Umrechnungsformel wie von Douglas O'Shaughnessy in [O'S87] verwendet.

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \tag{14}$$

Kompression mithilfe des Logarithmus

Im nächsten Schritt wird auf $F[k_{mel}]$ Gleichung 15 angewendet.

$$\hat{F}[k_{mel}] = \log(cF[k_{mel}] + 1) \tag{15}$$

Die Anwendung des Logarithmus beruht auf der Idee des Quelle-Filter Modells (Sektion 2.1.1). Der Logarithmus (Gleichung 15) soll helfen $e[n]$ und $h[n]$ (Gleichung 2) zu trennen. Durch eine Trennung kann man eine möglichst explizite Betrachtung von $h[n]$ also der Klangcharakteristik erreichen. Anregungssignal und Filter sind im Zeitbereich durch Faltung und im Frequenzbereich multiplikativ verknüpft. Durch die Logarithmierung im Frequenzbereich soll die multiplikative in eine additive Verknüpfung umgewandelt werden (Gleichung 16).

$$\log(EH) = \log(E) + \log(H) \tag{16}$$

Additiv verknüpfte Signale sind schließlich leichter aufzutrennen. Da Gleichung 15 keine reine Anwendung des Logarithmus ist, ist zu beachten dass die Umwandlung von multiplikativer zu additiver Verknüpfung nur approximiert wird.
Die Addition von eins in Gleichung 15 ist deshalb von Nöten da für spätere Berechnungen sehr große negative Werte $\hat{F} \ll 0$ vermieden werden sollen. Diese Addition führt allerdings zu unerwünschten Nichtlinearitäten, welche durch einen Skalierungsfaktor c möglichst gering gehalten werden sollen. Für eine großen Skalierungsfaktor c ist wiederum zu berücksichtigen, dass die Gefahr besteht, dass dieser F überlagert [SG09].

Anwendung der Diskreten Kosinus Transformation (DCT)

Im letzten Schritt wird für $\hat{F}[k_{mel}]$ die DCT durchgeführt.

$$mfcc[k_{mel}] = \sum_{n=1}^{N_{mel}} \hat{F}[k_{mel}] \cos\left(\frac{\pi(n-0.5)k_{mel}}{N_{mel}}\right) \tag{17}$$

Durch die DCT können die relevanten Informationen extrahiert werden. Aufgrund der DCT können nämlich in der Praxis die hohen MFCC fallen gelassen werden. Durch das Fallenlassen wird die Trennung von $e[n]$ und $h[n]$ approximiert, da $h[n]$ vor allem den tiefen MFCC zugeordnet werden kann. In Abbildung 6 wird die Wirkung des Fallenlassens auf ein Spektrum veranschaulicht.

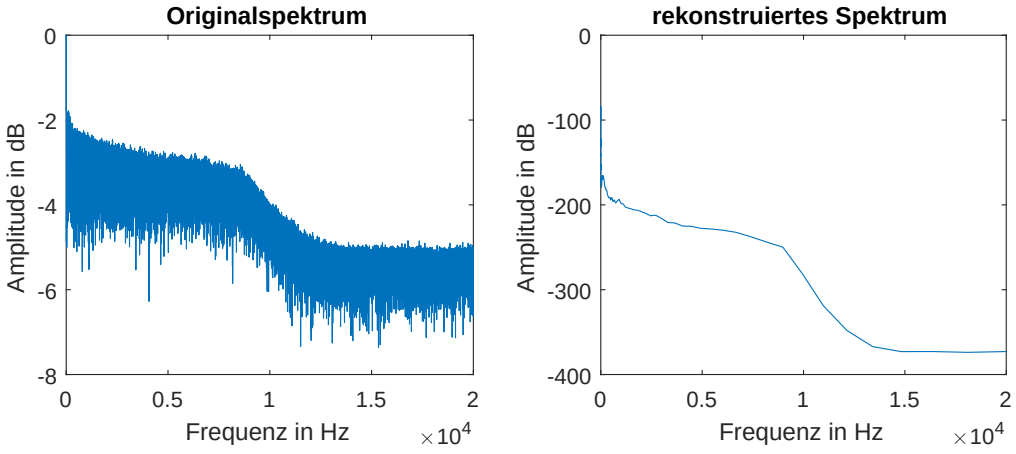

Abbildung 6 – Das linke Bild zeigt das Originalspektrum eines Beispielmusiksignals. Für das rechte Spektrum wurde die Betrags-DFT kosinustransformiert und eine Rekonstruktion ohne die hohen DCT-Koeffizienten durchgeführt. Es ist sichtbar, dass das Fallenlassen der hohen DCT-Koeffizienten die Welligkeit des Spektrums eliminiert. Es wird relevante Information über die Klangcharakteristik extrahiert. Die Amplitude ist rechts geringer, da durch das Fallenlassen der hohen MFCC, Signalenergie verloren geht.

2.1.5 K-Means

Die allgemeine Theorie des k-means Algorithmus ist in [M$^+$67] erläutert. Die hier verwendete Variante ist in [SG09] zu finden. K-means ist ein rekursiver Algorithmus zur Gruppierung von Daten. Es wird von einem Datenset $P = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_L\}$ ausgegangen, wobei $\mathbf{p}_l$ einen N-dimensionalen Vektor darstellt. Es ist das Ziel die L Vektoren in K Gruppen $A = \{A_1, A_2, ..., A_K\}$ einzuteilen. Jede Gruppe A_k enthält eine beliebige Anzahl J_k an Vektoren.
Vor Beginn des Algorithmus wird das Datenset statistisch standartisiert. Dabei wird von jedem Vektor der Mittelwert abgezogen und es wird die Varianz auf eins skaliert.

$$\mathbf{p}_l \leftarrow \mathbf{p}_l - \frac{1}{L}\sum_{l=1}^{L}\mathbf{p}_l \tag{18}$$

$$\mathbf{p}_l \leftarrow \mathbf{p}_l \frac{1}{\sqrt{\sum_{l=1}^{L}\mathbf{p}_l^2}} \tag{19}$$

Im ersten Schritt werden dann die Vektoren $\{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_L\}$ zufällig den Gruppen $\{A_1, A_2, ..., A_k\}$ zugeordnet. Danach beginnt der rekursive Teil des Algorithmus. Es wird für jede Gruppe das arithmetische Mittel berechnet. Wodurch für jede der K Gruppen ein zugehöriger Mittelwertvektor $\boldsymbol{\mu}_k$ der Dimension N resultiert.

$$\boldsymbol{\mu}_k = \frac{1}{J_k} \sum_{A_k} \mathbf{p}_l \tag{20}$$

Auf Basis der Mittelwerte werden die Gruppen neu geordnet. Der Vektor $\mathbf{p}_l$ wird der Gruppe zugeordnet, bei der die euklidische Distanz (Gleichung 21) zum zugehörigen Mittelwertvektor am geringsten ist.

$$d(\boldsymbol{\mu}_k, \mathbf{p}_l) = \sqrt{\sum_{n=1}^{N} (\mathbf{p}_l - \boldsymbol{\mu}_k)^2} \tag{21}$$

Die beiden Schritte (Mittelwertberechnung und Zuordnung) werden so oft wiederholt bis die Zuordnung von Iteration zu Iteration gleich bleibt, oder ein obere Schwelle an Iterationen erreicht wird.
Der Algorithmus konvergiert aufgrund der anfänglich zufälligen Gruppierung nicht immer zum gleichen Ergebnis. Die Qualität der Ergebnisse kann variieren. Um die Qualität verschiedener Durchführungen zu vergleichen kann die Summe der Gruppenvarianzen σ^2 gebildet werden.
Es ist anzumerken, dass alternativ zur initialen Zuordnung der Gruppen, auch eine zufällige Wahl der Mittelwertvektoren möglich ist. Damit würde k-means, durch die immer gleiche Wahl der Startwerte, immer zum gleichen Ergebnis führen. Allerdings entspricht dieses Ergebnis dann nicht unbedingt der Zuordnung mit der kleinsten Varianz.

$$\sigma^2 = \sum_{k=1}^{K} \sigma_k^2 \tag{22}$$

$$\sigma_k^2 = \sum_{n=1}^{N} \sum_{A_k} (\mathbf{p}_l - \boldsymbol{\mu}_k)^2 \tag{23}$$

Um den Algorithmus zu optimieren, können mehrere Durchführungen für das selbe Datenset gemacht werden und daraufhin das Ergebnis gewählt werden, bei dem σ^2 am geringsten ist.

2.2 Vorgehensweiße und prozeduale Schritte

Der Algorithmus zur Quellseparation wird in 5 Schritten beschrieben:

1. Transformation in die Zeit-Frequenzebene

2. Zerlegung des Signals in Komponenten
3. Extraktion der Information über die Klangcharakteristik
4. Gruppierung der Komponenten in Quell-Gruppen
5. Synthese der Quellen

Die Ausgangssituation ist eine monaurale Signalmischung $x[n]$ vor, welche aus verschiedenen Quellen $s_m[n]$ zusammengesetzt ist (Gleichung 1).

Transformation in die Zeit-Frequenzebene

$$x[n] \rightarrow \underline{\mathbf{X}} \rightarrow \mathbf{X} \tag{24}$$

Im ersten Schritt wird eine STFT mit der Signalmischung durchgeführt (Sektion 2.1.2). Das Signal wird somit im Zeit-Frequenzbereich durch eine komplexwertige Matrix $\underline{\mathbf{X}}$ dargestellt. Daraufhin wird von jedem Wert der Matrix $\underline{\mathbf{X}}$ der Betrag gebildet. Es ergibt sich die Matrix $\mathbf{X}$.

Zerlegung des Signals in Komponenten

$$\mathbf{X} \rightarrow \{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_I\} \tag{25}$$

Die Quellseparation funktioniert auf Basis der Amplitudenwerte. Phasenwerte werden nicht herangezogen. Dementsprechend wird im nächsten Schritt eine NMF (Sektion 2.1.3) mit $\mathbf{X}$ durchgeführt. Es ergeben sich I Komponenten. Jede Komponente $\mathbf{C}_i$ besteht aus einem Frequenzspektrum $\mathbf{B}_i$ und dessen zeitlichem Verlauf $\mathbf{G}_i$.

Extraktion der Information über die Klangcharakteristik

$$\{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_I\} \rightarrow \{\mathbf{B}_1, \mathbf{B}_2, ..., \mathbf{B}_I\} \rightarrow \{\mathbf{mfcc}_1, \mathbf{mfcc}_2, ..., \mathbf{mfcc}_I\} \tag{26}$$

Die Signalkomponenten sollen später in Quellen gruppiert werden. Da die Gruppierung auf Basis von Klangcharakteristiken geschehen soll, muss die relevante Information über die Klangcharakteristik aus $\mathbf{C}_i$ extrahiert werden. Es werden folglich nicht die gesamten Matrizen $\{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_I\}$ herangezogen, sondern nur die korrespondierenden Frequenzspektren $\{\mathbf{B}_1, \mathbf{B}_2, ..., \mathbf{B}_I\}$. Um die Gruppierung zu verbessern, werden die Frequenzspektren zu MFCC transformiert (Sektion 2.1.4). Es werden die hohen MFCC fallengelassen. Außerdem wird auch der erste MFCC fallengelassen, da er hauptsächlich die Signalamplitude repräsentiert [NSHAG07]. Eine Gruppierung mit MFCC eignet sich besser als eine Gruppierung mit Frequenzspektren. Das liegt daran, dass die MFCC die Information über die Klangcharakteristik komprimiert beinhalten und überflüssige Information durch die Transformation extrahiert wird.

Gruppierung der Komponenten in Quell-Gruppen

$$\{\mathbf{mfcc}_1, \mathbf{mfcc}_2, ..., \mathbf{mfcc}_I\} \rightarrow \{A_1, A_2, ..., A_M\} \rightarrow \{A'_1, A'_2, ..., A'_M\} \tag{27}$$

Die MFCC stellen das Datenset dar, mit welchem ein k-means Algorithmus (Sektion 2.1.5) durchgeführt wird: $\mathbf{p}_l = \mathbf{mfcc}_i$. Der Algorithmus wird für $K = M$ Gruppen durchgeführt. Die Anzahl der Gruppen K entspricht also der bekannten Anzahl der Quellen M. Nach erfolgreicher Durchführung sind die MFCC-Vektoren den Gruppen $\{A_1, A_2, ..., A_M\}$ zugeordnet. Jeder MFCC-Vektor $\mathbf{mfcc}_i$ korrespondiert zu einer Komponente $\mathbf{C}_i$. Für die Gruppen $\{A'_1, A'_2, ..., A'_M\}$ wird jeder Vektor $\mathbf{mfcc}_i$ aus $\{A_1, A_2, ..., A_M\}$ durch seine korrespondierende Komponente $\mathbf{C}_i$ ersetzt. Letztendlich erhält man die Gruppen $\{A'_1, A'_2, ..., A'_M\}$, wobei die Gruppe A'_m die Komponenten $\{\mathbf{C}_i, ...\}$ eines Quellsignals $s_m[n]$ enthalten soll.

Synthese der Quellen

$$\{A'_1, A'_2, ..., A'_M\} \rightarrow \{\tilde{\mathbf{S}}_1, \tilde{\mathbf{S}}_2, ..., \tilde{\mathbf{S}}_M\} \rightarrow \{\tilde{s}_1[n], \tilde{s}_2[n], ..., \tilde{s}_M[n]\} \tag{28}$$

Für die Synthese der Quellen werden alle Komponenten $\{\mathbf{C}_i, ...\}$ einer Quellgruppe A'_m aufsummiert.

$$\tilde{\mathbf{S}}_m = \sum_{A'_m} \mathbf{C}_i \tag{29}$$

$\tilde{\mathbf{S}}_m$ ist eine Matrix von Betragsspektren. Für die Rücktransformation in den Zeitbereich benötigt man allerdings auch die Phasenlage. Man kann eine komplexe STFT-Matrix $\underline{\tilde{\mathbf{S}}}_m$ konstruieren indem man $\tilde{\mathbf{S}}_m$ als Maske auf die Matrix $\underline{\mathbf{X}}$ anwendet. $\tilde{\mathbf{S}}_m$ muss hierfür mit $\tilde{\mathbf{X}} = \mathbf{BG} = \sum_{i=1}^{I} \mathbf{C}_i$ normiert werden.

$$\underline{\tilde{\mathbf{S}}}_m[n,k] = \underline{\mathbf{X}}[n,k] \frac{\tilde{\mathbf{S}}_m[n,k]}{\tilde{\mathbf{X}}[n,k]} \tag{30}$$

Zum Schluss kann mit $\underline{\tilde{\mathbf{S}}}_m$ eine inverse STFT durchgeführt werden um ein approximiertes Quellsignal $\tilde{s}_m[n]$ zu erhalten

$$\tilde{s}_m[n] = ISTFT\{\underline{\tilde{\mathbf{S}}}_m\} \tag{31}$$

3 Evaluierung der Parametrierbarkeit

3.1 Ziele der Evaluierung

Es treten im Algorithmus zur Quellseparation mehrere variablen Größen (zum Beispiel: I, N_{mel}, ...) auf. Diese variablen Größen werden als Parameter verstanden. Das Ziel ist

es zu untersuchen wie sich der Algorithmus durch Veränderung dieser Parameter verhält. Inwieweit durch das Verändern der Parameter die Ergebnisse steuerbar und optimierbar sind, wird durch die Parametrierbarkeit beschrieben.
Im Folgenden werden folgende Aspekte der Parametrierbarkeit untersucht:

1. Grenzen der Parametrierbarkeit durch Ergebnisschwankungen
2. Allgemeine Voreinstellung der Parameter

Grenzen der Parametrierbarkeit durch Ergebnisschwankungen

Im Laufe des Algorithmus finden zweimal Zufallsinitialisierungen statt. Einmal werden zu Beginn der NMF $\mathbf{B}$ und $\mathbf{G}$ zufällig initialisiert. Zusätzlich basiert die erste Gruppenzuordnung im k-means auf Zufall. Das führt dazu, dass die Quellseparation nicht deterministisch erfolgt. Bei mehrfacher Durchführung mit derselben Signalmischung $x[n]$ ergeben sich unterschiedliche synthetisierte Quellen. Die Qualität, also die Ähnlichkeit zu den Originalquellen, variiert auch. Daraus folgt, dass, da die Ergebnisse mit jeder Durchführung schwanken, der Algorithmus nur in Grenzen parametrierbar ist. Eine Untersuchung der Ergebnisschwankungen ist notwendig, um die Grenzen der Parametrierbarkeit zu erfassen.

Allgemeine Voreinstellung der Parameter

Der Algorithmus geht von einem blinden Ansatz zur Quellseparation aus. Dementsprechend nimmt man an, dass keine Information über die Quellen vorliegt. Das Ziel ist es deshalb die Parameter so zu optimieren, dass sie unabhängig von den Quellen zu guten Resultaten führen. Um den Arbeitsaufwand nicht zu groß werden zu lassen, wird hauptsächlich untersucht, in welcher Größenordnung sich die Parameter bewegen sollten.

3.2 Evaluierungsprozess

Ausgangssituation

Für die Evaluierung werden sechs Originalquellen aus dem EBU-Datensatz [EBU08] verwendet (siehe Tabelle 1).

Track-Nummer von [EBU08]	Inhalt	Zeitausschnitte
08	Violine	0:30-0:40
26	Claves	0:17-0:27
39	Klavier	1:50-2:00
49	weibliche Stimme (Englisch)	0:00-0:10
52	männliche Stimme (Französisch)	0:00-0:10
58	Gitarre	0:00-0:10

Tabelle 1 – Auflistung der Originalquellen welche aus [EBU08] stammen

Da es es sich um Stereoaufnahmen handelt, wurde jeweils der linke (erste) Kanal dieser Aufnahmen verwendet. Bei den Instrumenten wurden die Zeitausschnitte so gewählt, dass

Musik beziehungsweise Sprache an der gewählten Stelle vorhanden ist. Zusätzlich wird ein Fade-in und ein Fade-out von $0.5s$ eingefügt. Die Amplituden werden jeweils mit 0.5 multipliziert um eine Übersteuerung beim mischen zu verhindern.
Für die Evaluierung werden Signalmischungen aus jeweils zwei Quellen erzeugt. Somit liegen $\binom{6}{2} = 15$ Signalmischungen vor.

Evaluierungsmethode

Für die Evaluierung wurde die Software von [EVHH11] und [Vin12] genutzt. Die Software vergleicht eine separierte Quelle mit der Originalquelle. Im Rahmen der Softwareerstellung wurde ein Hörversuch mit 23 Experten durchgeführt. Daraufhin wurden Zusammenhänge zwischen den Signalen der Testquellen und den Bewertungen der Experten untersucht. Das Ziel der Software ist es eine Expertenbewertung einer separierten Testquelle vorauszusagen.
Die Testquellen wurden anhand von vier Aspekten evaluiert (siehe dazu Tabelle 2). Die Software retourniert für jeden Aspekt einen Score zwischen 0 und 100, wobei ein hoher Wert eine gute Quellseparation im Bezug auf diesen Aspekt bedeutet.
Da man für jede Separation in dieser Arbeit von zwei Originalquellen ausgeht und man

Evaluierungsaspekt	Evaluierungsscores
Auftreten von Artefakten	APS (Artifacts-related Perceptual Score)
Auftreten von Interferenzen mit anderen Quellen	IPS (Interference-related Perceptual Score)
Verzerrung der Originalquelle	TPS (Target-related Perceptual Score)
insgesamte Bewertung	OPS (Overall Perceptual Score)

Tabelle 2 – Evaluierungsaspekte aus [EVHH11] und die zugehörigen Punktestände

dementsprechend zwei separierte Quellen erhält, wird über die zwei resultierenden Scores gemittelt um für eine Separation einen Score zu erhalten.

Identifizierung der Quellen

Da der Quellseparationsalgorithmus die Quellen nicht identifizieren kann, muss dies anhand der Evaluierungsscores geschehen. Bei jeder Quellseparation erhält man zwei separierte Quellen, die mit jeweils zwei Originalquellen verglichen werden. Man berechnet folglich vier mal die Scores. Für die Identifizierung wird der insgesamt höchste OPS dieser vier Scores genutzt. Die diesbezügliche Korrespondenz aus separierter und entsprechender Originalquelle stellt die erste Identifizierung dar. Die übrig bleibende separierte Quelle wird als die andere Quelle identifiziert.

3.3 Evaluierung der Ergebnisschwankungen

Für die Evaluierung der Ergebnisschwankungen wurde die Quellseparation der 15 Signalmischungen 10 mal durchgeführt. Daraufhin wurden die Scores berechnet. Die Parameter wurden wie in Tabelle 3 dargestellt, gewählt.

Über die 10 Durchführung wurden wurden Mittelwerte gebildet. Zudem wurden aus den

I	25
N_{mel} (vor Fallenlassen der Mel-Bänder)	20
N_{mel} (nach Fallenlassen der Mel-Bänder)	9
c	1
maximale Iterationen bei der NMF	100
maximale Iterationen bei k-means	100

Tabelle 3 – Parametereinstellung bei der Analyse der Ergebnisschwankungen

10 Durchführungen die Standardabweichungen, Minima und Maxima berechnet. Die Resultate sind in den Tabellen 4 und 5 zu sehen.

	μ				σ			
Signalmischung	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS
Claves und weibliche Stimme	22.8	57.1	31.7	49.6	1.9	14.9	3.3	9.0
Claves und Gitarre	18.5	48.0	29.2	49.3	0.2	1.5	3.1	1.0
Claves und männliche Stimme	26.2	47.8	34.4	53.4	3.6	12.8	0.9	7.6
Claves und Klavier	20.0	87.1	22.3	59.5	1.8	3.4	2.2	4.8
Claves und Violine	14.3	35.7	15.8	39.8	2.1	19.1	2.6	9.8
weibliche Stimme und Gitarre	19.3	44.4	23.4	52.7	2.3	7.8	5.4	4.7
weibliche Stimme und männliche Stimme	17.4	43.1	22.9	52.6	2.0	1.6	8.1	5.5
weibliche Stimme und Klavier	12.9	26.4	11.5	44.6	1.3	11.5	4.8	4.6
weibliche Stimme und Violine	11.6	19.3	10.6	32.2	2.4	7.0	3.6	6.2
Gitarre und männliche Stimme	20.7	40.1	21.8	52.0	1.3	9.7	2.6	7.4
Gitarre und Klavier	12.6	37.9	21.4	60.3	3.7	3.3	2.6	5.6
Gitarre und Violine	9.7	19.7	8.3	28.5	1.2	8.4	2.1	6.8
männliche Stimme und Klavier	16.6	43.2	23.2	59.2	3.4	9.0	2.4	3.2
männliche Stimme und Violine	15.6	21.8	12.7	30.0	1.4	7.4	3.9	8.8
Klavier und Violine	11.7	54.1	15.9	53.8	3.0	7.6	2.4	1.9

Tabelle 4 – arithmetische Mittel (links) und Standardabweichungen (rechts) der 10 Durchführungen

In welchem Bereich die Punktestände schwanken kann in Tabelle 5 eingesehen werden. Aus Tabelle 4 ist ersichtlich dass die Scores stark von der Signalmischung abhängig sind. Zudem variiert die Größenordnung der Standardabweichung ebenfalls mit den verschiedenen Signalmischungen.
Beispielsweise schwankt der OPS für die Mischung Gitarre und Claves nur mit einer Standardabweichung von 0.2 Punkten, während sich für die Mischung von männlicher Stimme und Claves 3.6 Punkte ergeben. Anhand von Tabelle 6 ist ersichtlich, wie konstant welcher Sorce ist. Die Daten deuten darauf hin, dass mit steigendem Score die Schwankungen tendenziell größer werden.

	Min				Max			
Signalmischung	OPS	TPS	IPS	APS	OPS	TPS	IPS	APS
Claves und weibliche Stimme	20.6	28.2	28.3	34.4	24.6	76.3	35.8	57.6
Claves und Gitarre	18.2	46.2	25.0	48.3	18.7	49.7	33.3	50.7
Claves und männliche Stimme	21.3	27.3	33.6	46.8	31.6	67.0	36.7	64.4
Claves und Klavier	18.2	81.2	19.0	63.7	24.4	91.9	25.0	75.1
Claves und Violine	11.7	24.9	12.6	32.4	17.0	92.4	21.3	69.0
weibliche Stimme und Gitarre	16.0	32.7	15.3	45.6	21.9	50.4	28.4	56.9
weibliche Stimme und männliche Stimme	14.4	41.3	11.9	45.3	19.4	45.5	32.7	59.2
weibliche Stimme und Klavier	11.6	20.8	8.5	42.7	16.0	60.8	25.4	58.3
weibliche Stimme und Violine	8.6	9.7	5.3	23.2	14.3	27.1	13.9	40.6
Gitarre und männliche Stimme	16.8	25.6	14.8	30.0	21.3	54.6	24.8	56.2
Gitarre und Klavier	8.7	33.3	17.9	52.9	17.1	43.2	24.5	69.9
Gitarre und Violine	8.3	9.9	6.0	17.7	11.0	31.8	11.4	34.0
männliche Stimme und Klavier	13.3	33.4	17.4	52.6	21.7	57.0	25.1	62.0
männliche Stimme und Violine	14.4	15.1	10.1	21.1	18.3	35.2	20.3	44.9
Klavier und Violine	8.6	45.1	11.8	49.9	16.1	63.5	18.2	56.1

Tabelle 5 – Minima (links) und Maxima (rechts) der 10 Durchführungen

	μ
OPS	16.7
TPS	41.7
IPS	20.3
APS	48.5

	σ
OPS	2.1
TPS	8.3
IPS	3.3
APS	5.7

Tabelle 6 – durchschnittliche Scores (links) und Standardabweichungen (rechts); gemittelt über die 10 Durchführungen und alle Signalmischungen

3.4 Evaluierung von Voreinstellungen

Für die Evaluierung der Voreinstellungen der Parameter wurden weitere Quellseparationen mit unterschiedlichen Parametereinstellungen durchgeführt. Die Daten aus Sektion 3.3 konnten als die erste Voreinstellung verwendet werden. Zusätzlich wurden für I und N_{mel} je 4 weitere Werte getestet. Die anderen Parameter blieben dabei wie in Tabelle 3. Es wurden die Quellseparationen für jede Parametereinstellung drei mal durchgeführt um den Einfluss der Ergebnisschwankungen zu verringern. Die Resultate der drei Durchführungen wurden gemittelt.

Parameter I

Der Parameter I gibt an, in wie viele Komponenten die NMF die Signalmischung aufspaltet (siehe Sektion 2.1.3). Es wurden diesbezüglich die folgenden fünf Werte untersucht: $5, 12, 25, 50, 100$.

In Abbildung 7 kann man erkennen, dass der Einfluss des Parameters I über alle Signalmischungen gemittelt relativ gering bleibt. Die Differenz zwischen den geringstem und höchstem Score liegt für alle Scores unter 10 Punkten. Es ist ein leichter Anstieg zu den niedrigeren Werten sichtbar (vor allem bei OPS und IPS). Dementsprechend lässt

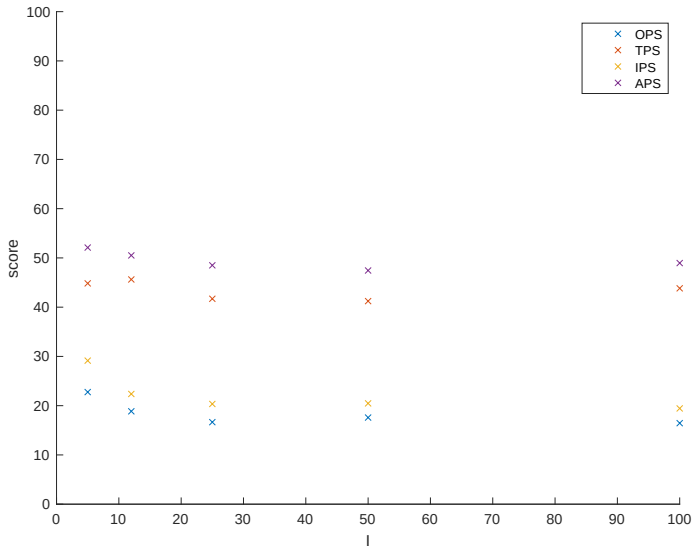

Abbildung 7 – gemittelte Scores für die unterschiedlichen Werte für $I = 5, 12, 25, 50, 100$

sich eine Optimierbarkeit durch geringere Werte für I vermuten. Es liegt vor allem die Vermutung nahe, dass geringe Werte von I, Interferenzen verringern können. Um einen allgemeinen Zusammenhang zwischen dem Parameter und einer guten Quellseparation sicher zu stellen, ist die Datenlage aber unzureichend.

Parameter N_{mel}

N_{mel} gibt an, wie viele MFCC verwendet werden (siehe Sektion 2.1.4). Es ist zu beachten, dass circa die Hälfte der MFCC fallengelassen werden. Für die Evaluierung wurden vor dem Fallenlassen der Mel-Bänder die Werte $N_{mel} = 6, 10, 20, 40, 80$ getestet. Nach dem Fallenlassen bleiben bei den Tests $N_{mel} = 3, 5, 9, 18, 36$ MFCC übrig.

Abbildung 8 zeigt, dass der Einfluss von N_{mel} auch relativ gering bleibt. Die Werte variieren so wie bei I mit weniger als 10 Punkten. APS und TPS korrelieren für die Ergebnisse. IPS und OPS korrelieren ebenfalls. Während APS und TPS für hohe MFCC gute Ergebnisse liefern, sind IPS und OPS für geringere Werte besser. Aufgrund einer teilweisen Gegenläufigkeit lässt es sich schwer beurteilen was eine gute Wahl für die Anzahl der MFCC ist.

4 Diskussion und Schlussfolgerungen

Diskussion des Algorithmus

Der untersuchte Algorithmus ist nicht derjenige, der zum jetzigen Zeitpunkt die besten Ergebnisse für die Quellseparation liefert. Allerdings beinhaltet er Elemente (zum Beispiel: NMF, MFCC), welche nach wie vor eine Rolle für den blinden Ansatz zur Quellseparati-

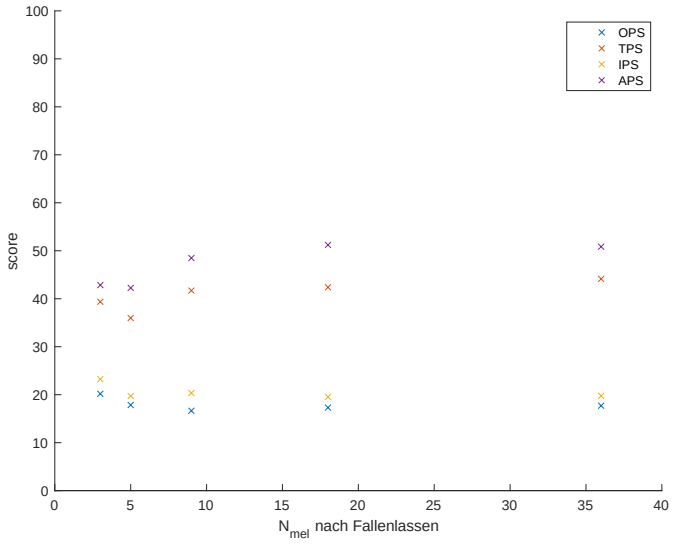

Abbildung 8 – gemittelte Scores für die unterschiedlichen Werte für $N_{mel} = 3, 5, 9, 18, 36$

on spielen.
Ob für einen blinden Ansatz zur Quellseparation das Quelle-Filter-Modell zutrifft, ist fraglich. Das Modell geht schließlich grundsätzlich von einer möglichen Trennung von Filter und Quelle aus (siehe Sektion 2.1.1). Bei vielen Klängen entspricht dieses Modell allerdings nicht der Realität. So entstehen Klangcharakteristiken oft auch durch die ursprüngliche Klangerzeugung. Das Quelle-Filter-Modell wurde im Kontext der MFCC verwendet. Hier zeigt jedoch [CP18], dass das Modell zumindest für Musik durchaus eine Berechtigung hat. So konnten bei dieser Arbeit Instrumentenklänge zu 90% richtig klassifiziert werden.

Diskussion der Evaluierungsergebnisse

Die Menge der generierten Daten zur Evaluierung ist zu gering um Tatsachen zu schlussfolgern. Aufgrund der hohen Rechendauer der Evaluierungssoftware, war eine größere Datenmenge im Rahmen der Bachelor-Arbeit nicht möglich. Um eine größere Aussagekraft zu erzielen, sollten mehr Signalmischungen verwendet werden und die Scores sollten über mehr Durchführungen gemittelt werden.
Außerdem wurden einige möglichen Einflüsse nicht untersucht. So hat die Pegeldifferenz der zu separierenden Quellen einen Einfluss auf die Separation [SG09]. Darüber hinaus wurde nicht berücksichtigt, dass die Parameter sich gegenseitig beeinflussen könnten. Außerdem wurden auch nicht alle Parameter (siehe Tabelle 3) untersucht. Letztlich wäre es auch sinnvoll noch mehr Werte für die Parameter zu untersuchen.
Um die Relevanz der Ergebnisse zu beurteilen muss berücksichtigt werden, dass die Scores nur versuchen eine Expertenbewertung vorauszusagen. Dabei liegt die beispielsweise für den OPS die Präzision bei 90% und die Monotonie bei 76% [EVHH11].
Da die Differenzen der Evaluierungscores unter 10 Punkten liegen, die generierte Datenmenge nicht groß ist und die Evaluierungsmethode eine Unsicherheit hat, kann keine allgemein gültige Aussage über eine optimale Voreinstellung der untersuchten Parameter

getroffen werden. Man kann allerdings feststellen, dass der Einfluss der Parameter gering bleibt (unter 10 Punkten für die untersuchten Werte).

Literatur

[CP18] S. S. Chakraborty and R. Parekh, "Improved musical instrument classification using cepstral coefficients and neural networks," in *Methodologies and Application Issues of Contemporary Computing Framework*. Springer, 2018, pp. 123–138.

[DM80] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[EBU08] EBU. (2008) Sund quality assessment material aecordings for subjective tests. [Online]. Available: https://tech.ebu.ch/publications/sqamcd

[EVHH11] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[Fan70] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970, no. 2.

[HKV09] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation." in *ISMIR*, 2009, pp. 327–332.

[KL08] K. Kokkinakis and P. C. Loizou, "Using blind source separation techniques to improve speech recognition in bilateral cochlear implant patients," *The Journal of the Acoustical Society of America*, vol. 123, no. 4, pp. 2379–2390, 2008.

[M$^+$67] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[ME07] M. I. Mandel and D. P. Ellis, "Em localization and separation using interaural level and phase cues," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2007, pp. 275–278.

[NSHAG07] A. B. Nielsen, S. Sigurdsson, L. K. Hansen, and J. Arenas-García, "On the relevance of spectral features for instrument classification," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 2. IEEE, 2007, pp. II–485.

[O'S87] D. O'Shaughnessy, "Speech communication, human and machine addison wesley," *Reading MA*, 1987.

[PJLL99] H.-M. Park, H.-Y. Jung, T.-W. Lee, and S.-Y. Lee, "Subband-based blind signal separation for noisy speech recognition," *Electronics Letters*, vol. 35, no. 23, pp. 2011–2012, 1999.

[SG09] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proceedings of the 12th International Conference on Digital Audio Effects*, 2009.

[Sma04] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2004, pp. 494–499.

[Vin12] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 430–437.

[Vir07] T. Virtanen, "Monoaural sound source seperation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 1066–1074, 2007.