Verbesserte Singing Voice Separation für Riddim-Alben

Jens Johannsmeier[1], Kenneth Allan[1], Sebastian Stober[1]

[1] *Otto-von-Guericke Universität, 39106 Magdeburg, Deutschland, Email: {jens.johannsmeier, kenneth.allan, stober}@ovgu.de*

Einleitung

Instrumentalversionen von Musikstücken haben vielseitige Anwendungen. Dazu zählen zum Beispiel Karaoke oder das Remixen zum Erstellen alternativer Versionen. Allerdings sind Instrumentalversionen nur selten frei zugänglich.

Das Ziel der Singing Voice Separation ist es, ein vorliegendes Musikstück in Gesangs- und Instrumentalspuren zu trennen. Somit können (theoretisch) Instrumentalversionen eines jeden beliebigen Stückes generiert werden. In den letzten Jahren wurden mächtige Open-Source-Werkzeuge veröffentlicht, die diese Aufgabe mit hoher Qualität bewerkstelligen können. Doch auch diese Methoden verursachen hörbare Artefakte in den Ergebnissen und sind somit verbesserungsfähig.

Eine besondere Situation ergibt sich für sogenannte Riddim-Alben aus dem jamaikanischen Reggae [1]. Hier singen verschiedene Künstler Texte auf denselben „Riddim“, was wir vereinfacht als Instrumental verstehen können. Unsere Hypothese ist, dass in dieser Situation eine verbesserte Wiederherstellung der Instrumentalversion möglich ist. Weiterhin enthalten viele Riddim-Alben das Original-Instrumental, was einen Vergleich der Ergebnisse mit dem Optimum ermöglicht.

Wir stellen im Folgenden eine Methode vor, welche die vorhergehenden Überlegungen umsetzt, und zeigen mittels quantitativen Metriken als auch einer qualitativen Online-Studie, dass wir die Ergebnisse einiger frei verfügbarer Systeme zur Klangquellentrennung signifikant verbessern können.

Hintergründe

Wir geben einen kurzen Überblick über bestehende Ansätze zur Klangquellentrennung. Außerdem verweisen wir kurz auf die Vordergrund-Hintergrund-Trennung in Videos, welche unserem Ansatz konzeptionell ähnelt.

Klangquellentrennung

Klassische Ansätze zur Klangquellentrennung basieren oft auf Annahmen wie statistisch unabhängigen Quellen und Aufnahmen mit mehreren Mikrofonen [2]. Im musikalischen Kontext sind diese Voraussetzungen in der Regel nicht erfüllt. Daher kommen hier andere Techniken zum Einsatz, wie etwa Nichtnegative Matrixfaktorisierung oder Harmonisch-Perkussiv-Trennung [3].

Aktuell erreichen vor allem Techniken aus dem Deep Learning [4] starke Ergebnisse. Diese werden „end-to-end“ trainiert, ein gegebenes Stück in vorher definierte Spuren zu trennen (z.B. Gesang, Begleitung, Schlagzeug...). Das Trainieren dieser Modelle erfordert große Datenmengen (in der Regel inklusive der bereits getrennten Spuren zur Berechnung der Fehlerfunktion) und hohe Rechenkapazitäten. Ferner ist die Qualität der Ergebnisse oft abhängig von der Ähnlichkeit des Stücks zu den Trainingsdaten – somit leidet vor allem die Trennung von unterrepräsentierten Musikstilen häufiger unter Artefakten. In dieser Arbeit verwenden wir zwei Deep-Learning-Methoden: Demucs [5] und Spleeter [6].

Vordergrund-Hintergrund-Trennung in Videos

In der Bildverarbeitung existieren Methoden, um einen variablen Vordergrund (z.B. eine sich bewegende Person) von einem statischen Hintergrund zu trennen. Unter milden Voraussetzungen kann beispielsweise der Median über die Zeit verwendet werden. Dies ist in Abbildung 1 dargestellt.

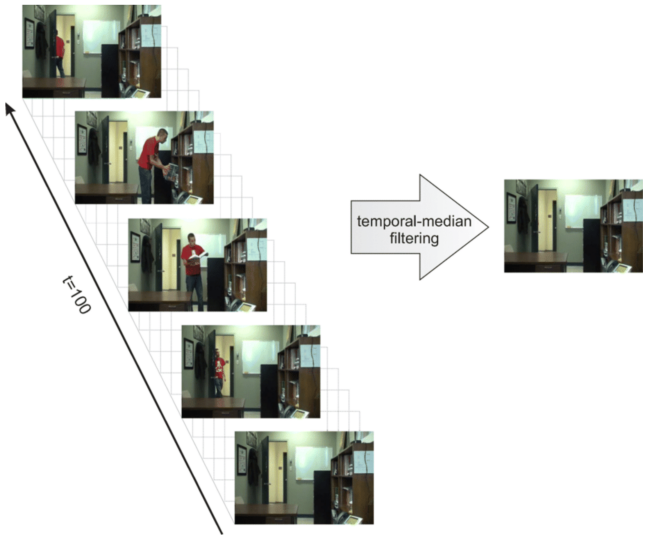

Abbildung 1: Medianbasierter Ansatz zur Vordergrund-Hintergrund-Trennung in Videos. Reproduziert aus [7].

Konzept

Wir beschreiben zwei Methoden zur Trennung von Stimme und Instrumental für Riddim-Alben. Die erste operiert direkt auf den vorliegenden Stücken eines Albums, wogegen die zweite zuerst ein bestehendes System auf die einzelnen Stücke anwendet. Ansonsten sind beide Methoden identisch. Abbildung 2 zeigt eine Übersicht.

Vorverarbeitung

Die Stücke werden bezüglich ihrer Lautstärker normalisiert. Anschließend wird eine Short-time Fourier-

Transformation (STFT) [3] angewendet, wobei wir eine Fenstergröße von 1024 Samples und Schrittweite von 128 Samples verwenden, bei einer Sampling Rate von 44100 Hz. In den nachfolgenden Schritten wird nur die Magnitude des Spektrogramms verwendet, nicht die Phase.

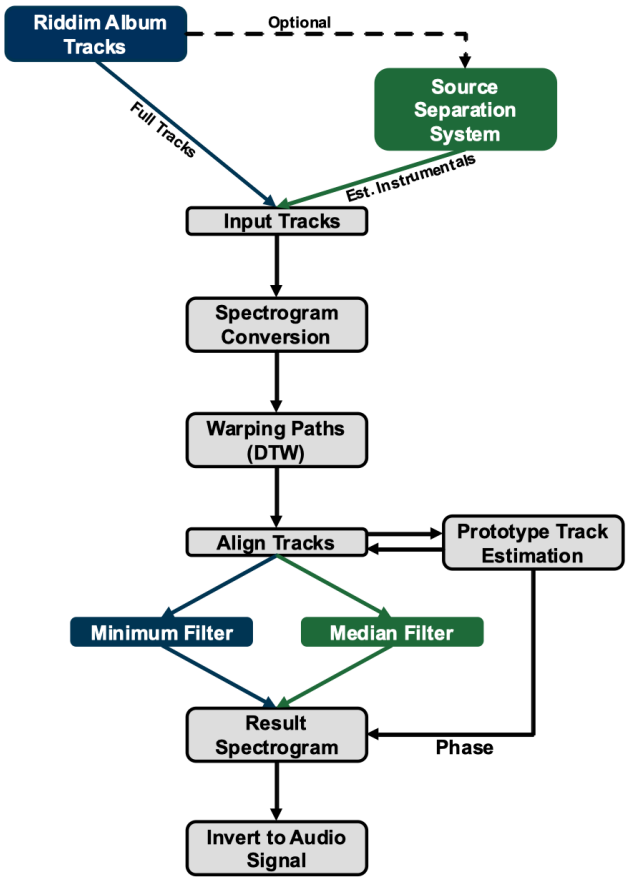

Abbildung 2: Übersicht des in dieser Arbeit verwendeten Systems.

Synchronisierung

Die Spektrogramme der Stücke eines Albums sollen daraufhin gefiltert werden. Hierfür müssen sie zeitlich synchronisiert werden. Dies wird dadurch erschwert, dass viele Stücke ein individuelles „Intro“ enthalten, welches vor dem Beginn des Instrumentals eingefügt wird. Auch können Stücke Tempovariationen enthalten – solche schließen wir allerdings aus der vorliegenden Studie aus.

Die Synchronisierung verwendet ein modifiziertes Dynamic Time Warping (DTW) [3] – da wir annehmen, dass die Stücke dasselbe Tempo haben und sich im Wesentlichen durch das Hinzufügen von Intros unterscheiden, sollte beispielsweise der „Hauptpfad“ eine Steigung von 1 haben. Dies kann vorgegeben oder durch ein nachfolgendes Smoothing approximiert werden. Im Zuge der Synchronisierung wird auch ein sogenannter Prototyp bestimmt – dies ist ein Stück, welches möglichst kein zusätzliches Intro im Vergleich zum echten Instrumental hat. Dass ein solches Stück existiert, wird hier angenommen. Eine ausführlichere Beschreibung der Methode ist aus Platzgründen hier nicht möglich.

Filtern

Anhand der synchronisierten Stücke wird das Instrumental herausgefiltert. Hierfür werden die jeweils synchronen Zeitschritte der einzelnen Stücke extrahiert und aggregiert. Die Aggregation erfolgt unabhängig für jedes Zeit- und Frequenzfenster als auch für die beiden Stereokanäle. Als Aggregationsfunktion wird das Minimum verwendet, wenn Stücke direkt als Eingabe fungieren, während bei vorgeschalteter Trennung durch Demucs oder Spleeter der Median verwendet wird. Die Begründung ist in Abbildung 3 dargestellt: Wenn Gesang vorhanden ist, sollte möglichst wenig spektrale Information beibehalten werden; bei den bereits getrennten Stücken hingegen sollten Artefakte jeglicher Art beseitigt werden – hier sind Ausreißer in beide Richtungen möglich.

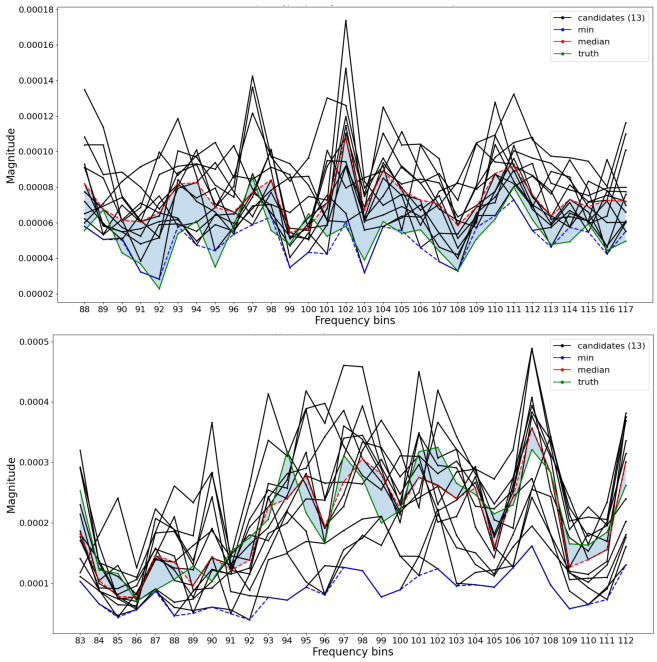

Abbildung 3: Magnituden für einige Frequenzbins, dargestellt für Stücke eines Riddim-Albums. Oben: Bei nicht getrennten Stücken liegt das Minimum nahe an den Werten des wahren Instrumentals. Unten: Bei bereits getrennten Stücken liegt der Median nahe an den wahren Werten.

STFT-Inversion

Zuletzt muss das erhaltene Spektrogramm zurück in den Audioraum transformiert werden. Für einen natürlichen Klang wird dabei Phaseninformation benötigt. Diese mit ähnlichen Methoden wie die Magnitude aus den gegebenen Stücken zu extrahieren, lieferte allerdings verrauschte Ergebnisse. Methoden zur Phasenschätzung wie Griffin-Lim [8] verursachen ebenfalls hörbare Artefakte. In dieser Arbeit verwenden wir daher die Phase des Prototypen für das abgetrennte Instrumental. Die gegebene Magnitude und Phase werden mittels der inversen STFT transformiert.

Tabelle 1: Die in dieser Arbeit verwendeten Riddims mitsamt einiger Metadaten.

Riddim-Album	Veröffentlichung	Anzahl nutzbarer Stücke	BPM
Aida	2019	4	85
Answer 2021	2021	3	80
Borderline	2006	12	70
Clappas	2003	22	160
Doctor's Darling	2004	16	70
Gully Slime	2006	3	93
Hardtimes	2007	14	77
Kuff Again!	2013	6	85
Martial Arts	2002	10	94
Prisonbreak	2009	12	75
Retro Teng	2019	5	84
Showtime	1997	6	90
The Sensi-millionaire	2011	17	78
Zion	2006	11	80

Experimente

Das vorgestellte System wurde auf einer Auswahl von geeigneten Riddim-Alben getestet. Eine Übersicht befindet sich in Tabelle 1.

Quantitative Evaluation

Wir berechnen die durchschnittliche Distanz zwischen den Resultaten der jeweiligen Methoden und der Ground Truth. Als Distanzmaße verwenden wir scale-invariant signal-to-distortion ratio (SI-SDR) [9] und die L2-Norm zwischen den Magnituden-Spektrogrammen. Die Ergebnisse sind in Tabelle 2 dargestellt. Die Benennung der Methoden ist wie folgt:

- `result_tracks`: Direktes Filtern der Stücke.
- `reference_spleeter`: Spleeter angewandt auf den Prototypen.
- `result_spleeter`: Trennung durch Spleeter und anschließendes Filtern.
- `reference_demucs`: Demucs angewandt auf den Prototypen.
- `result_demucs`: Trennung durch Demucs und anschließendes Filtern.

Es ist zu sehen, dass die direkte Anwendung von Spleeter oder Demucs auf den Prototypen bessere Ergebnisse erzielt als die direkte Anwendung unserer Methode. Allerdings führt unsere Methode, angewandt auf bereits getrennte Stücke, zu einer Verbesserung der Metriken, mit Ausnahme des SI-SDR für Demucs. Aufgrund der hohen Standardfehler ist es allerdings schwierig, hier klare Aussagen zu treffen. Bezüglich der relativen Qualität von Spleeter und Demucs ist kein eindeutiges Muster zu erkennen.

Tabelle 2: Mittlere quantitative Metriken der verschiedenen Methoden. In Klammern ist der empirische Standardfehler angegeben.

Variante	Metrik	
	L2 Spektrogramm	SI-SDR
result_tracks	67529.49 (17010.80)	-5.41 (1.88)
reference_spleeter	51888.05 (18147.51)	-4.72 (2.18)
result_spleeter	47545.77 (13515.68)	-4.60 (2.01)
reference_demucs	59444.12 (18182.72)	-4.71 (2.13)
result_demucs	40658.67 (12845.37)	-4.81 (2.03)

Qualitative Evaluation

Um einen Eindruck von der subjektiv wahrgenommenen Qualität der untersuchten Ansätze zu bekommen, führen wir eine Online-Nutzerstudie durch. Die Teilnehmer hören die verschiedenen Varianten der getrennten Riddims und bringen sie in eine Reihenfolge abhängig von der Qualität und dem Nicht-Vorhandensein von Gesang bzw. menschlichen Stimmen. Ferner können sie Kommentare zu den jeweiligen Varianten hinterlassen. Wir stellen Audiobeispiele zur Anhörung bereit.[1]

Insgesamt führten 65 Teilnehmer 248 Rankings durch. Eine Zusammenfassung der Ergebnisse findet sich in Abbildung 4 und Tabelle 3. Es ist ersichtlich, dass unsere Methode die Anwendung von Spleeter auf den Prototypen übertrifft, nicht aber Demucs. Ferner führt eine vorherige Trennung der einzelnen Stücke durch besagte Systeme zu einer deutlichen Qualitätssteigerung. Darüber hinaus wird Demucs durchgehend besser als Spleeter bewertet. Unsere Methode mit vorheriger Trennung durch Demucs wird ähnlich gut bewertet wie die Ground Truth, in einigen Fällen sogar besser. Auffällig ist, dass dies vor allem in Gruppen ohne musikalische Erfahrung oder mit auditorischen oder neurologischen Einschränkungen der Fall ist.

Signifikanzanalyse

Aus Platzgründen können wir hier keine ausführliche Analyse der statistischen Signifikanz der präsentierten Ergebnisse darstellen. Friedman- und Nemenyi-Tests [10, 11] zeigen allerdings, dass die quantitativen Metriken sich nicht bzw. nur in Ausnahmefällen unterscheiden, während die qualitativen Ergebnisse signifikant unterschiedlich sind.

[1] `https://figshare.com/articles/media/Riddim_Extraction_Results/19629996`

Tabelle 3: Mittlere subjektive Rankings der verschiedenen Methoden. In Klammern ist der empirische Standardfehler angegeben.

Variante	Teilnehmer				
	Alle	Musikalische Erfahrung	Keine Erfahrung	Einschränkungen	Keine Einschränkungen
ground_truth	2.34 (0.095)	2.16 (0.091)	2.55 (0.098)	2.72 (0.106)	2.30 (0.093)
result_tracks	4.41 (0.095)	4.34 (0.095)	4.50 (0.096)	3.96 (0.089)	4.46 (0.096)
reference_spleeter	5.0 (0.083)	5.32 (0.059)	4.61 (0.100)	5.16 (0.075)	4.98 (0.085)
result_spleeter	3.11 (0.087)	3.17 (0.082)	3.04 (0.093)	3.00 (0.091)	3.12 (0.087)
reference_demucs	3.78 (0.088)	3.76 (0.086)	3.81 (0.091)	4.00 (0.090)	3.76 (0.088)
result_demucs	2.36 (0.080)	2.26 (0.075)	2.49 (0.085)	2.16 (0.087)	2.39 (0.079)

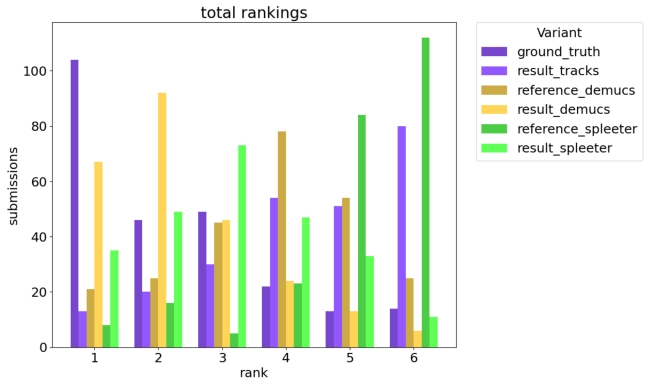

Abbildung 4: Ranking-Histogramme der einzelnen Methoden. Für jede Methode ist aufgetragen, wie häufig diese mit einem bestimmten Rang bewertet wurde.

Zusammenfassung

Wir haben ein System vorgestellt, welches sich das Vorhandensein mehrerer Musikstücke mit demselben bzw. ähnlichem Hintergrund zunutze macht, um die Abtrennung des Gesangs bzw. Wiederherstellung des Instrumentals zu verbessern. Eine qualitative Nutzerstudie als auch quantitative Metriken zeigen, dass unsere Methode die Ergebnisse, welche durch Systeme wie Spleeter oder Demucs erreicht werden, deutlich verbessern kann.

Insgesamt zeigt unsere Arbeit, dass Klangquellentrennung durch zunutze machen domänenspezifischer Eigenschaften der betroffenen Stücke verbessert werden kann. Das hier dargestellte System ist zwar sehr spezifisch auf Riddim-Alben ausgelegt; es ist allerdings davon auszugehen, dass auch in anderen Musikstilen spezifische Eigenschaften ausgemacht werden können, die zu einer Verbesserung der Ergebnisse führen. Die vorliegende Studie soll hierfür als Beispiel dienen.

Literatur

[1] Lamarche Manuel, P., Marshall, W.: The riddim method: aesthetics, practice, and ownership in Jamaican dancehall. Popular Music 25 (2006), 447–470

[2] Comon, P., Jutten, C.: Handbook of Blind Source Separation: Independent Component Analysis and Applications. Academic Press, 2010

[3] Müller, M.: Fundamentals of Music Processing — Audio, Analysis, Algorithms, Applications. Springer Verlag, 2015

[4] LeCun, Y., Bengio, Y., Hinton, G.: Deep Learning. nature 521.7553 (2015), 436-444

[5] Défossez, A., Usunier, N., Bottou, L., Bach, F.: Music source separation in the waveform domain. arXiv preprint (2019), arXiv:1911.13254

[6] Hennequin, R., Khlif, A., Voituret, F., Moussallam, M.: Spleeter: a fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software, 5.50 (2020), 2154

[7] Cordova-Esparza, D.-M., Terven J., Jiménez-Hernández, H., Herrera-Navarro, A., Vázquez Cervantes, A., Garcia Huerta, J. M.: Low- bandwidth 3D visual telepresence system. Multimedia Tools and Applications 78 (2019)

[8] Griffin, D., Lim, J.: Signal estimation from modified short-time Fourier transform. IEEE Transactions on acoustics, speech, and signal processing, 32.2 (1984), 236-243

[9] Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R.: SDR – Half-baked or Well Done? IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019), 626–630

[10] Friedman, W.: Friedman Test. The SAGE Encyclopedia of Research Design (2022)

[11] Nemenyi, P. B.: Distribution-free multiple comparisons. Princeton University, 1963