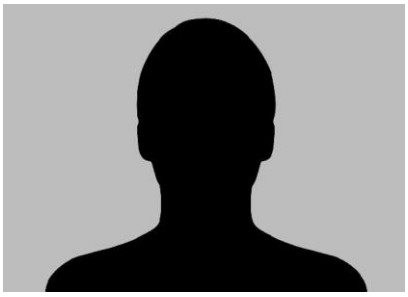
Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

Neural Network-Based Audio Source Separation



Diplomierende
Flavian Burkhard
Luca Neukom

Dozierende
Martin Loeser
Matthias Rosenthal

Das Ziel dieser Arbeit ist es, verschiedene Instrumente durch einen Blind Source Separation (BSS) Ansatz mit neuronalen Netzwerken aus einem Mix zu separieren. Sie baut auf einer Vorgängerarbeit auf, die sich auf Gesangsspuren konzentrierte. In dieser Arbeit werden nun verschiedene Ansätze untersucht, um melodische Instrumenten aus einem Musik-Mix zu extrahieren.

Diese Arbeit konzentriert sich auf die Instrumente Klavier und Gitarre. Es wurde festgestellt, dass einige Ideen aus der Vorgängerarbeit zwar anwendbar scheinen, aber nicht ohne Anpassungen auf diese Instrumente angewendet werden können. Aufgrund von Einschränkungen der öffentlich verfügbaren Datensätze musste ein neuer Datensatz von Gitarren- und Klavierspuren erstellt werden, mit denen die Netze trainiert werden konnten. Audiospuren mussten gesammelt, kategorisiert und in zwei Mixes gemischt werden, eine mit nur dem Zielinstrument und eine mit allen anderen Instrumenten. Der neue Datensatz besteht aus MedleyDB, MedleyDB 2.0 sowie Tracks aus den Spielen Rockband und Guitar Hero sowie der Mixing-Webseite Cambridge Music Technology.

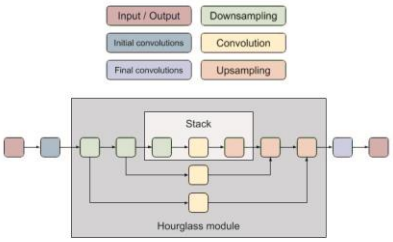
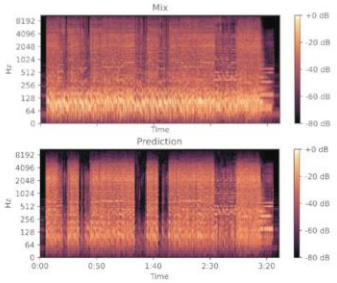


Diagramm der Stacked Hourglass Netzwerkstruktur mit einem Hourglass-Modul bestehend aus drei Stacks.

Es wurde eine Netzwerkstruktur namens Stacked Hourglass verwendet. Der Hauptunterschied zwischen den Trainings, die im Rahmen dieser Arbeit durchgeführt wurden, ist der Umfang der Vorverarbeitung der Trainingsdaten, die Grösse der Datensätze und das Verhältnis von gemischten Datensätzen. Die Auswertung erfolgte mit mir_eval und den drei gängigen Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR) und Source-to-Artifact Ratio (SAR).

Der beste erreichte SDR-Wert war 0.73 für die Gitarre durch Verwendung eines möglichst grossen Trainingsdatensatzes. Für das Klavier wurde ein SDR-Wert von -7.28 erreicht. Für die besten Resultate war es erforderlich, die Trainingsdaten einem Preprocessing zu unterziehen. Ein Vergleich der Trainings auf den zwei Instrumenten mit ähnlichen Datensätzen zeigt, dass die Ergebnisse auch bei gleichen Parametern sehr unterschiedlich ausfallen. Im Allgemeinen werden jedoch mit einem grösseren Datensatz bessere Ergebnisse erzielt als mit anderen Versuchen die Trainingssätze zu verbessern.



Ein Spektrogramm eines vollen Mixes und ein Spektrogramm des separierten Gitarrenteils.

Ein simpler Web-Demonstrator wurde erstellt, um einen einfachen Zugriff auf die trainierten Modelle zu ermöglichen. Der gesamte Quellcode dieser Arbeit wurde auf Github veröffentlich.