Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering
InES Institute of Embedded Systems

Deep Vocal Extraction from Music

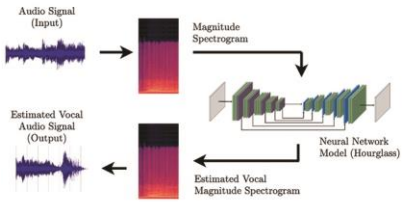
Diplomierende
David Flury
Andreas Kaufmann
Raphael Müller

Dozierende
Martin Loeser
Matthias Rosenthal

Das Ziel dieser Arbeit ist es, *Blind Audio Source Separation (BASS)* mithilfe von Neuronalen Netzwerken anzuwenden, um Musik in abgemischter Form in ihre Gesangs- und Instrumentalspuren zu teilen.

Es handelt sich um einen Bereich des Forschungsgebietes *Music Information Retrieval (MIR)*. *BASS* findet in vielen verschiedenen Bereichen wie zum Beispiel der Optimierung der Sprachqualität in Hörgeräten oder der Nachvertonung von Filmen Anwendung. Für Musikbegeisterte ist es interessant, um Musikstücke in ihre Gesangs- und Instrumentalspuren aufzuteilen.

Das Ziel ist erstens, frühere Ergebnisse reproduzieren zu können, zweitens eine Sammlung von Trainingsdaten zusammenzustellen, welche über 1'000 Lieder umfasst und damit existierenden Sammlungen sowohl in Qualität als auch Quantität überlegen ist und drittens, verschiedene auf das Problem der *BASS* anwendbare Architekturen neuronaler Netze zu untersuchen und zu vergleichen. Eine optimierte Netzwerkarchitektur wird vorgeschlagen. Zudem wird eine Webapplikation entwickelt, mithilfe derer Audiodateien in Instrumental und Gesang aufgeteilt werden können.



Aus dem Audiosignal eines Musikstücks werden Spektrogramme erzeugt. Das vorher trainierte Netzwerk nimmt diese Daten als Input und berechnet daraus eine Schätzung für das Spektrogramm der Gesangsspur, welche in ein Audiosignal zurückverwandelt wird.

Als Vorverarbeitung werden die Trainingsdaten, die als Rohdaten in Audiodateien vorliegen, in komplexe Spektrogramme umgewandelt. Das Netzwerk erhält die Magnituden dieser Spektrogramme als Input. Der Output entspricht den Magnituden-Spektrogrammen der kalkulierten Gesangs- und Instrumentalspur, welche wieder in ein Audiosignal umgewandelt werden. Untersucht werden in der Arbeit auch Arten, wie die Spektrogramme normalisiert werden können.

Die Qualität des Netzes wird automatisch bewertet. Um die Resultate mit Arbeiten anderer Gruppen vergleichen zu können, werden die *Source to Distortion Ratio (SDR)*, *Source to Inference Ratio (SIR)* und *Source to Artefacts Ratio (SAR)* evaluiert. Diese sind in *BSSEval v4* (*mir_eval*) implementiert.

Frühere Ansätze konnten sowohl reproduziert als auch übertroffen werden. Durch die Verwendung des erweiterten Trainingssets und des optimierten Netzwerkes konnte ein *SDR-Wert* von 8.1 für die Gesangsspur und 14.2 für die Instrumentalspur erzielt werden, was einer Verbesserung von rund 64 % beziehungsweise 31 % zu einem vorjährigen Ansatz entspricht. Die Werte entsprechen dem heutigen Stand der Forschung.

Webapplikation zur Trennung der Stimmen und Instrumente in Musik

Die voll funktionsfähige Web-Applikation ist unter www.unmix.io verfügbar und funktioniert mit verschiedenen Quellen wie Datei-Uploads oder Youtube-Links.