Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

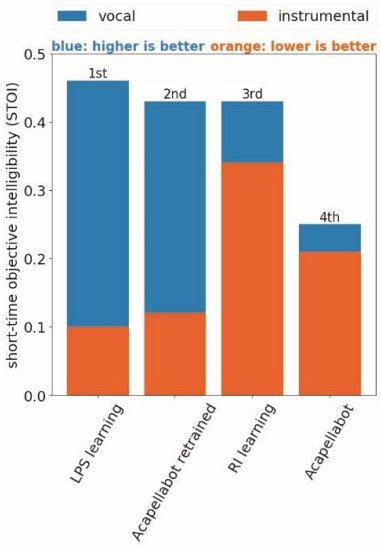
ISC Institute for Signal Processing
and Wireless Communications

Deep Convolutional Neural Network for Vocal Isolation from Music

Diplomierende
Raphael Freudiger
Fabian Strebel

Dozent
Martin Loeser

In dieser Arbeit soll die Leistung des AcapellaBots bei der Trennung von Gesang und Instrumenten in Liedern reproduziert werden. In einem weiteren Schritt sollen verschiedene Verbesserungsmöglichkeiten geprüft werden. Hierfür wird ein tiefes Convolutional Neural Network mit dem Datensatz der MedleyDB trainiert. Dieser beinhaltet sowohl separate Gesangsspuren und Instrumente sowie deren Mischungen, welche für das Training des neuronalen Netzes verwendet werden. Zur Verbesserung werden drei Methoden der Stichprobengenerierung sowie unterschiedliche Stichprobengrössen evaluiert. Zudem werden drei Modelle getestet (AcapellaBot retrained, Real and Imaginary, Log-power Spectrogram). Für den Leistungsvergleich der Modelle wird die Short-Term Objective Intelligibility als Mass für die Stimmverständlichkeit verwendet. Der ursprüngliche AcapellaBot erreicht eine Verständlichkeit von 0.25, was niedriger ist als jene der unverarbeiteten Mischung mit einem Wert von 0.39. Das erste Vergleichsmodell ist der neu trainierte AcapellaBot. Es verwendet eine andere Aktivierungsfunktion als der ursprüngliche AcapellaBot sowie die Daten der MedleyDB. Dieses Modell erreicht eine Verständlichkeit von 0.43, was eine deutliche Verbesserung gegenüber dem ursprünglichen AcapapellaBot darstellt. Das nächste, optimierte Modell, in welchem die Real- und Imaginärteile der komplexen Spektrogramme für das Training verwendet werden, erreicht ebenfalls eine Verständlichkeit von 0.43. Das letzte, optimierte Modell wird mit den Log-Power-Spektrogrammen trainiert. Die beste Leistung wird mit zufällig platzierten Stichroben mit einer Grösse von 256x256 erreicht. Die Phase geht in den Log-Power-Spektrogrammen verloren und wird am Ende durch sukzessive Approximation rekonstruiert. Trotz des Phasenverlusts erzielt dieses Modell schlussendlich die beste Stimmverständlichkeit von 0.46. Die Leistung des AcapellaBots konnte reproduziert werden. Durch die Optimierung der Stichprobengenerierung und Stichprobengrösse konnte die Leistung zudem fast verdoppelt werden.



Die vier Ansätze wurden mit dem STOI bewertet. Ein höherer Gesangs-STOI bedeutet eine bessere Verständlichkeit. Der ursprüngliche AcapellaBot hat die geringste Verständlichkeit. Es folgt das RI-Lernen und das nochmals trainierte AcapellaBot-Modell. Das RI-Lernen erreicht den gleichen Gesangs-STOI wie der nochmals trainierte AcapellaBot, hat aber einen schlechteren Instrumenten-STOI. Das LPS-Lernen mit optimierten Parametern hat den besten Gesangs- und Instrumenten-STOI.