Zurich University
of Applied Sciences

School of Engineering

InES Institute of Embedded Systems

Vocal Extraction from Music using Deep Learning

Diplomierende
David Flury
Andreas Kaufmann
Raphael Müller

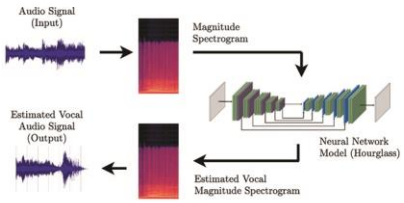
Dozierende
Martin Loeser
Matthias Rosenthal

The aim of this work is to apply*Blind Audio Source Separation (BASS)* using neural networks, in particular to separate music audio signals into voice and instrumental parts.

The challenge is part of the research field *Music Information Retrieval (MIR).* Apart from the usefulness of audio source separation on its own (for instance for musicians), *BASS* has applications in many other areas such as optimizing voice quality in hearing aids or post-production of audio in movies.

The goal is to first reproduce previous results, second to compile a suitable training set that is superior to existing ones both in terms of quality and quantity, and thirdly to examine and compare different neural networks architectures applicable to the problem of *BASS*. An optimized network architecture is proposed. In addition, a web application is developed which allows users to split audio files into their instrumental and vocal parts.

For pre-processing, the songs which are available as raw audio files are converted into complex spectrograms. The network receives the magnitudes of these spectrograms as input. The output corresponds to the magnitude spectrograms of the estimated vocal and instrumental tracks, which are converted back into an audio signal. The work also examines ways in which the spectrograms can be normalized.



Spectrograms are generated from the audio signal of a piece of music. The previously trained network takes this data as input and calculates an estimate for the spectrogram of the voice track, which is converted back into an audio signal.

The quality of the network is evaluated automatically. In order to be able to compare the results with those of other work, the *Source to Distortion Ratio (SDR)*, *Source to Inference Ratio (SIR)* and *Source to Artefacts Ratio (SAR)* are calculated. These are implemented in *BSSEval v4* (*mir_eval*).

We found that earlier approaches could be reproduced and outperformed. By using the extended training set and the optimized network, the best results achieved are an *SDR* value of 8.1 for the vocal track and 14.2 for the instrumental track. This corresponds to an improvement of about 64 % and 31 %, respectively, compared to a previous year's approach. The values correspond to the state-of-the-art in current research.

Web application to separate vocal and instrumental tracks from music

The fully functional web application is available at **www.unmix.io** and works with various sources such as file uploads or Youtube links.