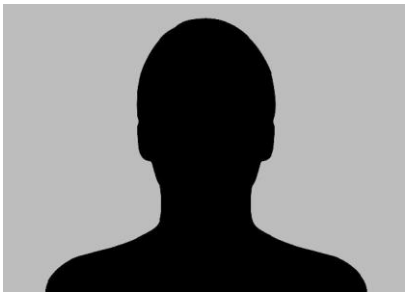
Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

ISC Institute for Signal Processing
and Wireless Communications

DNN based Speech Enhancement using Generative Adversarial Network Techniques



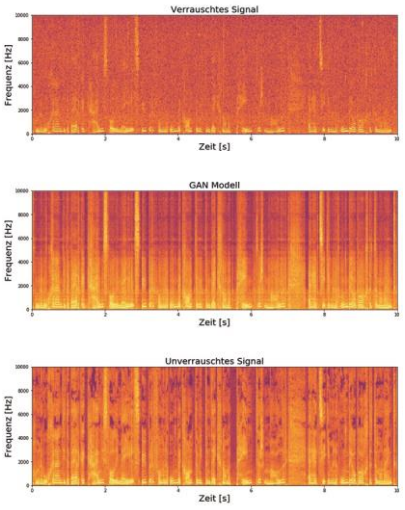
Diplomierende
Stefan Stajic
Stefan Wick

Dozent
Sigisbert Wyrsch

Das Verbessern der Verständlichkeit in Sprachsignalen (Speech Enhancement) ist ein hochaktuelles Forschungsgebiet mit vielen Anwendungsmöglichkeiten. In den letzten Jahren wurden in diesem Gebiet unter anderem Deep Neural Networks (DNN) verwendet, um die Sprachverständlichkeit zu verbessern. Dennoch weisen Audiosignale, die mit solchen DNN verarbeitet wurden, Verarbeitungsartefakte auf, die für Hörer störend wirken können. In dieser Bachelorarbeit sollen verschiedene DNN-Strukturen mit Modellen von sogenannten Generative Adversarial Networks (GANs) verbessert und miteinander verglichen werden. Eine typische Anwendung von einem GAN ist beispielsweise die Erzeugung von fotorealistischen Bildern oder natürlich wirkenden Stimmen. Im Gegensatz zu einer klassischen DNN-Struktur verwendet ein GAN vereinfacht erklärt zwei neuronale Netze (Generator und Diskriminator), die versuchen, sich gegenseitig zu trainieren.

Als Basis wurde die frequenzbasierte Speech-Enhancement-DNN-Struktur von Jean-Marc Valin verwendet. Dieses NN wurde zu einer GAN-Struktur erweitert. Des Weiteren wurden zum Vergleich weitere bestehende Speech-Enhancement-GAN-Strukturen implementiert. Eine zeitbasierte Autoencoder-Struktur, welche als solche kein GAN darstellt, wurde von Santiago Pascual et al. zu einem GAN erweitert (Speech Enhancement GAN, kurz: SEGAN). Deepak Baby erweiterte das SEGAN und entwickelte das Speech Enhancement Relativistic GAN (SERGAN) sowie das improved SEGAN (iSEGAN).

Für die Evaluation dieser Modelle wurden rauschfreie Interview-Ausschnitte aus dem Schweizer Radio und Fernsehen (SRF) mit vier verschiedenen Hintergrundgeräuschen mit unterschiedlichen SNR-Werten gemischt. Aus den Testergebnissen verschiedener objektiver Metriken stellte sich heraus, dass zeitbasierte Modelle das Rauschen mit tiefem SNR besser unterdrücken können als frequenzbasierte. Jedoch unterdrücken alle Modelle bei hohem SNR das Rauschen etwa gleich stark. Dennoch konnte durch die GAN-Erweiterung des Valin-Modells eine deutliche Verbesserung festgestellt werden. Das zeigt, dass frequenzbasierte Modelle, wie das Valin-Modell, durch eine GAN-Erweiterung und die Wahl der richtigen Trainingsparameter verbessert werden können.



Vergleich durch Spektrogramm: Oben das verrauschte Audiosignal, unten das unverrauschte Audiosignal und in der Mitte das durch das GAN-Modell gefilterte Audiosignal.