Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

ISC Institute for Signal Processing
and Wireless Communications

Robust Voice Activity Detection based on Statistical Models and Machine Learning

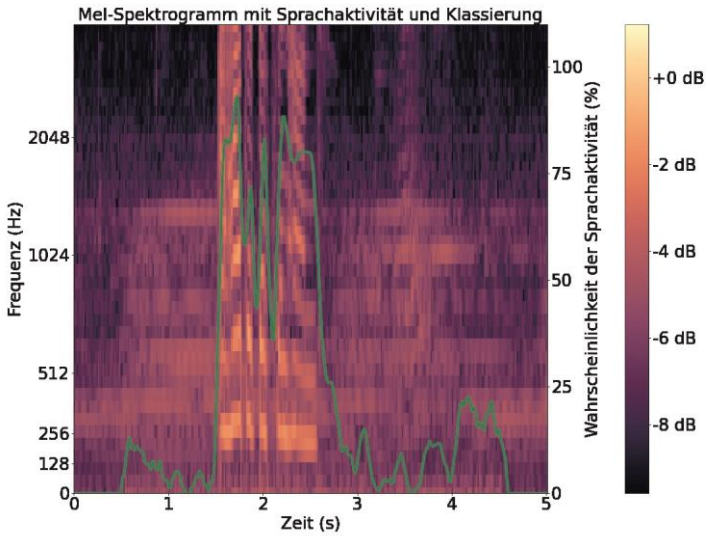
Diplomierende
Benaris Dizdarevic
Yannick Wälti

Dozierende
Philipp Matthias Schmid
Sigisbert Wyrsch

Diese Arbeit befasst sich mit der Entwicklung eines gegenüber Noise und Hall robusten Voice Activity Detector (VAD). Das Unterscheiden zwischen Sprache und Hintergrundgeräuschen ist eine wichtige Aufgabe in der Sprachsignalverarbeitung. Um diese Aufgabe zu lösen, wurden verschiedene neuronale Netzwerke trainiert, die als VAD eingesetzt werden können.

Aus einer Mischung der Sprachaufnahmen aus den VCTK und LibriSpeech Datensätzen mit dem Noise aus WHAM und Soundcities wurde ein Datensatz mit über 100 Stunden erstellt. Mit gezielter Augmentation und Verhallung wurde zusätzlich die Vielfalt des Datensatzes erhöht. Eine Auswahl von insgesamt fünf neuronalen Netzwerken mit Anwendung in VAD-Aufgaben wurden anschliessend durch Anpassung der Architektur, Hyperparameter und Feature-Daten verbessert und auf diesen Daten trainiert, bevor sie auf ihre Realtime-Fähigkeit und Klassierungsqualität verglichen wurden. Eine Anpassung der besten, Realtime-fähigen Netze in je drei Varianten mit unterschiedlicher Parameterzahl wurde ebenfalls evaluiert.

Die Resultate einer Evaluation mit den Testdaten aus den AVA und Kaist Datensätzen zeigen gute Resultate mit AUC-Werten im Bereich von 78 bis zu 98 %. Im Vergleich mit dem VAD von WebRTC zeigte sich, dass Ansätze mit neuronalen Netzen unabhängig von deren Grösse in allen Bereichen klar bessere Resultate erzielen. Dies gilt umso mehr dann, wenn die zu klassierende Sprache stark verrauscht ist.



Mel-Spektrogramm eines verrauschten Sprachbeispiels aus dem Datensatz, zusammen mit der Sprachaktivität und den entsprechenden Klassifizierungen. Mel-Spektrogramme werden als Eingabe-Features für die meisten der neuronalen Netze verwendet, die in dieser Arbeit evaluiert wurden. Die Sprachaktivität ist deutlich im Bereich zwischen 1.5 s und 3 s zu erkennen (Mel-Bins = 64, NFFT = 256, Hop-Länge = 128 mit Hann-Fenster).