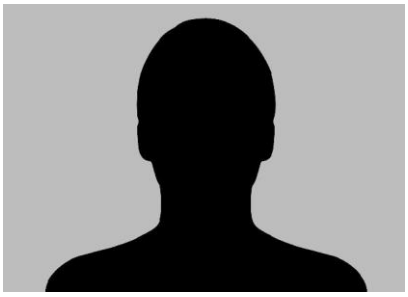
Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

InIT Institut für angewandte Informationstechnologie

Exploiting the Full Information of Varying-Length Utterances for DNN-Based Speaker Verification

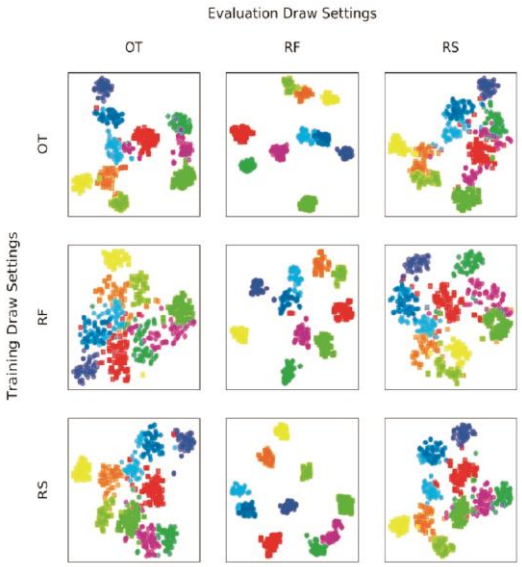


Diplomand/in
Daniel Neururer

Dozent/in
Thilo Stadelmann

Tiefe neuronale Netze wurden im Verlauf der letzten Jahre zum Standard in verschiedenen Bereichen der Sprecher-Erkennung. Obwohl in etlichen früheren Studien bereits versucht wurde, zeitabhängige Charakeristiken von Sprachaufnahmen mittels tiefen neuronalen Netzwerken zu modellieren, existieren bis jetzt noch keine Tests, die messen, wie gut sie speziell in diesem Aspekt sind.

Wir untersuchen verschiedene Architekturen in ihrer Fähigkeit, sich solche Zeitabhängigkeiten beim Modellieren eines Sprechers zu Nutzen zu machen. Hierfür schlagen wir einen neuen, aussagekräftigen Test vor, in welchem man die Leistung eines Netzwerks vergleicht, wenn man es mit beibehaltener zeitlicher Struktur bzw. mit randomisierter Zeitabhängigkeit in den Trainingsdaten trainiert und evaluiert. Zudem werden qualitative Analysen mittels verschiedener Visualisierungen der extrahierten Sprecher-Eigenschaften gemacht, welche wir mit den Erkenntnissen aus der Analyse der Leistung vergleichen. Unsere Tests deuten darauf hin, dass sogar renommierte Architekturen nicht Fähig dazu sind, klaren Nutzen aus beibehaltener Zeitabhängigkeit zu ziehen.



Modellierungen von 10 Sprechern, extrahiert durch die ResNet34S Architektur. Die mittlere Spalte visualisiert die extrahierten Sprecher-Modelle, wenn das Netzwerk Segmente mit randomisierter Zeit erhält.