Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

Voice Recognition with Deep Neural Networks

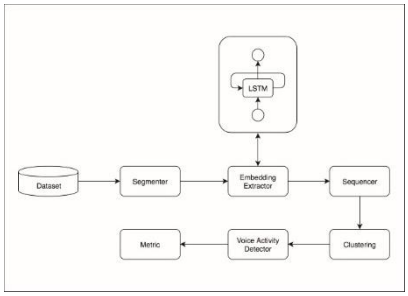
Eine typische Anwendung von *Stimmerkennung* ist die *Sprecherdiarisierung*. Für eine beliebige Tonaufnahme (z.B. ein Telefonanruf, eine Besprechungsaufnahme oder eine Tonspur eines TV-Programms) kann die Sprecherdiarisierung zusammen mit einem Sprache-zu-Text Modul verwendet werden, um die Tonquelle zu transkribieren. In diesem Fall würde das Sprache-zu-Text Modul den Text aus den gesprochenen Wörtern generieren, und das Sprecherdiarisierung-Modul für die Problemstellung 'wer wann gesprochen hat' zuständig sein.

In den vergangenen Jahren konnte das Institut für angewandte Informationstechnologie (InIT) grosse Erfolge im Bereich des Sprecherclustering verbuchen, einem Teil von Sprecherdiarisierung. Die vorliegende Arbeit versucht, diese neuen, auf maschinellem Lernen basierenden Ansätze zu vereinen und auf Sprecherdiarisierung anzuwenden. Dabei liegt der Fokus auf der Demonstration der Fähigkeiten dieser neuen Methoden sowie dem Aufdecken von bestehenden Lücken und Schwachstellen. Dies dient als Ausgangspunkt für ein Anknüpfen an die Leistung von State of the Art Sprecherdiarisierung-Systemen in der Zukunft.

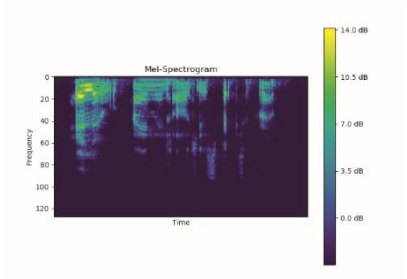
Verschiedene Experimente eruieren die Qualität der einzelnen Komponenten im System, um zukünftigen Projekten eine solide Basis mit Anknüpfungsmöglichkeiten zu bieten. Das entwickelte System ist fähig, im Setup der*NIST RT-09 evaluation campaign* anzutreten, um es mit anderen Sprecherdiariserung-Systemen zu vergleichen. Die gemessene Leistung zeigt ein durchgängig funktionstüchtiges, jedoch noch nicht wettbewerbsfähiges System.

Diplomierende
Niclas Simmler
Amin Trabi

Dozierende
Thilo Stadelmann
Oliver Dürr



Aufbau des in diesem Projekt entwickelten Sprecherdiarisierungs-systems.



Ein Mel-Spektrogramm ist eine visuelle Darstellung einer Tonaufnahme. Mel-Spektrogramme werden vom Sprecherdiarisierungssystem als Eingabe verwendet.