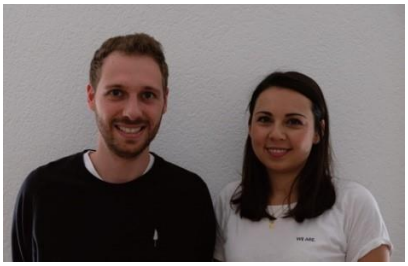
Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

InIT Institut für angewandte Informationstechnologie

Sprechererkennung mit wenigen Trainingsbeispielen



Diplomierende
Arik Sidney Guggenheim
Alicia Lea Rüegg

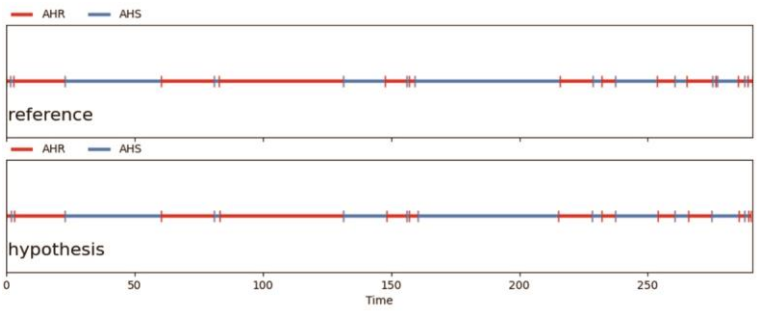
Dozent
Mark Cieliebak

Das Transkribieren von Interviews und Dialogen ist zeitaufwändig. Ein automatisiertes System könnte daher helfen, den Prozess zu beschleunigen. Bei der Entwicklung eines solchen Systems ist die Aufgabe der Sprechererkennung eine der grössten Herausforderungen.

In dieser Arbeit wird eine Lösung präsentiert, welche ein System zur Identifikation von Sprechern verwendet. Die Basis bildet ein Convolutional Neural Network, welches auf Audiodaten trainiert wurde. Dieses neuronale Netz wurde von der Visual Geometry Group entwickelt und trägt daher den Namen VGG.

Kurze, a-priori Stimmproben jedes Sprechers wurden verwendet, um das System dynamisch an ein neues Gespräch anzupassen. Heuristiken unterstützen bei der Interpretation der Ergebnisse. Der Sprachkorpus Verbmobil II wurde verwendet, um die optimalen Parameter unseres Systems zu ermitteln. Um einen Vergleich mit anderen aktuellen Sprecher-Erkennungs-Systemen zu ermöglichen, wurde die Verifikation mit dem LibriSpeech Korpus durchgeführt.

Es wurde festgestellt, dass Sprecherproben von 25 Sekunden ausreichen, um die Fehlerrate der Sprechererkennung (”DER”) zu minimieren. Auf dem Korpus LibriSpeech erreichte das System mit zwei Sprechern eine DER von 1.7% und eine Genauigkeit (”accuracy”) von 95.7%. Bei fünf Sprechern wurde eine DER von 4.8% und eine Genauigkeit von 91.8% erreicht.



Vergleich zwischen dem tatsächlichen und prognostizierten Sprecherwechsel eines Gesprächs. Die Referenz zeigt die tatsächlichen Segmente der Sprecher, während die Hypothese die Vorhersage durch das entwickelte System zeigt. Jeder der Sprecher wird durch eine andere Farbe dargestellt. Die auf diesem Gespräch berechnete Fehlerrate ("DER") beträgt 2.1%.