Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

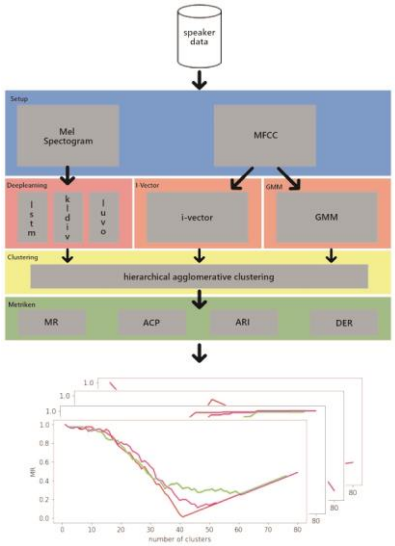
InIT Institut für angewandte Informationstechnologie

Benchmarking von klassischen und Deep Learning Speaker Clustering-Ansätzen

Diplomierende
Jan Sonderegger
Patrick Walter

Dozent
Thilo Stadelmann

Speaker Clustering beschreibt das Gruppieren von mehreren gesprochenen Segmenten nach den verschiedenen Sprechern, ohne die Anzahl der Sprecher oder die einzelnen Sprecher selbst zu kennen. In den vergangenen Jahren wurden dafür am Institut für angewandte Informationstechnologie der ZHAW mehrere Ansätze basierend auf Deep Learning entwickelt. Die erzielten Resultate dieser Ansätze waren vielversprechend, jedoch haben sich kleinere Unterschiede im Evaluationsprozess der jeweiligen Ansätze eingeschlichen. Aus diesem Grund waren die Resultate nicht komplett miteinander vergleichbar.
Das Ziel dieser Arbeit ist, die drei Ansätze 'Luvo', 'Pairwise_Kldiv' und 'Pairwise_Lstm' zu vergleichen. Um dies zu tun, müssen die Ansätze unter gleichen Voraussetzungen evaluiert werden, um den vielversprechendsten Ansatz bestimmen zu können. Zusätzlich sollen die Deep Learning-Ansätze mit klassischen Speaker Clustering-Ansätzen verglichen werden, um die erzielten Resultate besser einordnen zu können.
Um dieses Ziel zu erreichen, haben wir ein Benchmarking-System mit einem verbesserten Experimentaufbau entwickelt. Damit sollen vergleichbare und sinnvolle Resultate garantiert werden, indem wir bestehende Unterschiede in den Ansätzen eliminieren. Zusätzlich haben wir ein GMM- und ein I-Vector-System eingebaut. Die Resultate dieser beiden klassischen Speaker Clustering-Ansätze dienen als Referenzwerte, welche es zu übertreffen gilt. Um aussagekräftige Resultate zu bekommen, haben wir vier Metriken eingeführt, aus welchen sich verschiedene Rückschlüsse ziehen lassen.
Die Arbeit konnte die Deep Learning-Ansätze nicht abschliessend vergleichen. Die durchgeführten Experimente deuten an, dass die Ansätze optimiert werden müssen, um ihr Potential im neuen Experimentaufbau ausschöpfen zu können. Ausserdem hat sich gezeigt, dass der Pairwi-se_Lstm-Ansatz optimiert wurde, indem die Inputdaten in kleinere Segmente unterteilt werden. Dies könnte auch für Luvo und Pairwise_Kldiv eine lohnenswerte Verbesserung sein und müsste für einen aussagekräftigen Vergleich untersucht werden.



Dieses Bild zeigt die grundlegende Architektur unseres Systems. Es visualisiert die Speaker Clustering-Ansätze, welche in unserer Arbeit verglichen wurden, die dazugehörigen Vorverarbeitungsschritte für die Sprecherdaten sowie die vier Metriken, welche für den Vergleich der Resultate benutzt wurden. Die Plots unten im Bild zeigen, wie die drei Deep Learning-Algorithmen in unserem letzten Experiment abgeschnitten haben.