Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

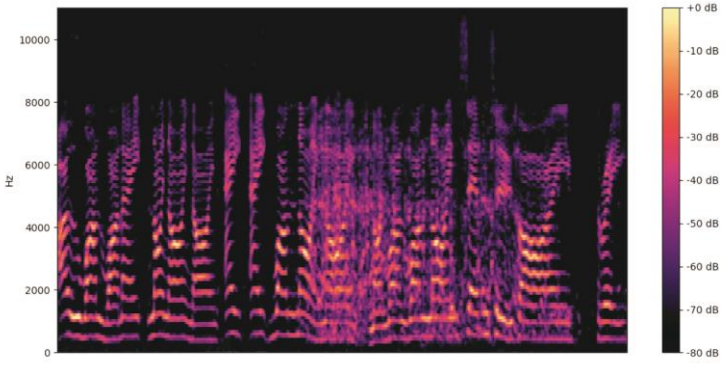
InIT Institut für angewandte Informationstechnologie

Speaker Clustering for Real-World Data using Deep Learning

Diplomierende
Christian Lauener
Claude Lehmann

Dozent
Thilo Stadelmann

In den letzten Jahren wurden in den Bereichen Speaker Recognition und Clustering durch die Verwendung von Deep Neural Networks beträchtliche Fortschritte erzielt. Durch die zeitabhängigen Aspekte von gesprochener Sprache werden Recurrent Neural Networks und Long Short-Term Memory-Strukturen auf eine Vielzahl von Problemen in diesem Bereich angewendet. Deren Fähigkeit, Entscheidungen auf zeitlich bereits zurückliegende Daten zu stützen, ist äusserst vorteilhaft für entsprechende Aufgaben.
Eine generelle Herausforderung beim Einsatz von Deep Learning ist, dass eine erhöhte Komplexität der Problemstellung meist eine grössere Menge an Trainingsdaten erfordert, um befriedigende Resultate zu erzielen. Üblicherweise wird dann zusätzliche Rechenleistung in Form von Hardware eingesetzt, um lange Trainingsphasen zu unterbinden. Dies ist etwa für kleinere Unternehmen oft kein gangbarer Weg.
In dieser Arbeit erforschen wir einen Active Learning-Ansatz, um die Datenmenge, welche ein neuronales Netzwerk zum Training benötigt, klein zu halten und dennoch akzeptable Resultate auf einem komplexen Datensatz zu erzielen. Wir präsentieren eine modulare Implementierung, mit welcher wir auf dem TIMIT-Datensatz mit 590 Sprechern einen neuen Bestwert erreichen, und zeigen auf, weshalb die Resultate auf dem VoxCeleb2-Datensatz nicht den Erwartungen entsprechen. Zusätzlich stellen wir Verbesserungen für das System vor, um es künftig weiter auszubauen.



Mit Hilfe von Spektrogrammen lässt sich Ton so darstellen, dass Bild-erkennungsalgorithmen darauf verwendet werden können. Wir benutzen neuronale Netze, um für jeden Sprecher Eigenschaften herauszukristallisieren, welche diesen eindeutig identifizierbar machen.
Da dies auf Tonaufnahmen mit Studioqualität bereits weitgehend gelöst ist, entwickeln wir einen bestehenden Ansatz weiter für Aufnahmen mit unterschiedlichsten Hintergrundgeräuschen, basierend auf derselben Datenrepräsentation.