Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

Automatische Stimmerkennung mit Deep Learning

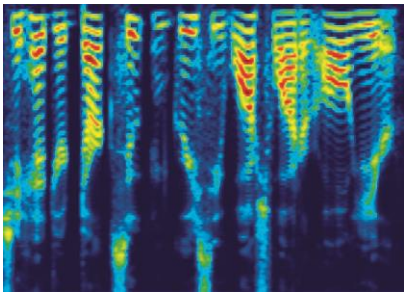
Diplomierende
Yanick Lukic
Carlo Vogt

Dozierende
Oliver Dürr
Thilo Stadelmann

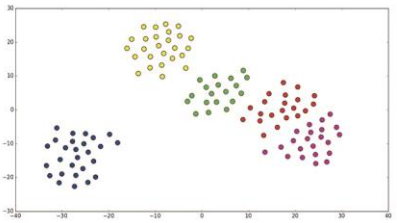
Deep Learning, besonders in Form von Convolutional Neural Networks (CNNs), hat in den letzten Jahren zu wesentlichen Verbesserungen in der Bildanalyse und verwandten Bereichen geführt. Dieser Fortschritt wird der Verschiebung weg von handgefertigten Features und individuellen Systemen hin zum automatisierten Lernen von Features aus nahezu unverarbeiteten Daten zugeschrieben. Für Sprecherclustering ist es jedoch noch immer üblich, handgefertigte Verarbeitungsketten wie MFCC Features und GMM-basierte Modelle einzusetzen.

In dieser Arbeit verwenden wir einfache Spektrogramme als Input für zwei unterschiedliche CNN Ansätze und evaluieren das optimale Design dieser Netzwerke für Sprecherclustering. Für den ersten Ansatz untersuchen wir die Frage, wie ein Netzwerk, das auf Sprecheridentifikation trainiert wurde, für Sprecherclustering eingesetzt werden kann. Der zweite Ansatz ist spezifisch für das Clustering von mehrdimensionalen Daten. Hierzu werden paarweise Constraints als Labels verwendet. Dazu verwenden wir eine mit der KL-Divergenz zusammenhängende Loss-Funktion, um die Distanz zwischen gleichen Paaren zu minimieren und zwischen ungleichen Paaren zu maximieren. Beide Ansätze verwenden das Konzept von Speaker Embeddings: Aktivierungen von Hidden-Layers eines vortrainierten Netzwerks stellen die Featurevektoren für das Clustering.

Wir demonstrieren unsere Ansätze auf dem bekannten TIMIT Datensatz und erreichen eine Misclassification Rate von 0.05 für das Clustern von 40 unbekannten Sprechern. State of the Art Ansätze erreichen bei identischem experimentellen Aufbau eine Misclassification Rate von 0.0625. Somit verbessern wir den State of the Art um 20.



Spektrogramme werden als Input für das neuronale Netzwerk verwendet.



t-SNE Darstellung der Featurevektoren, die für das Clustering verwendet werden.