Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

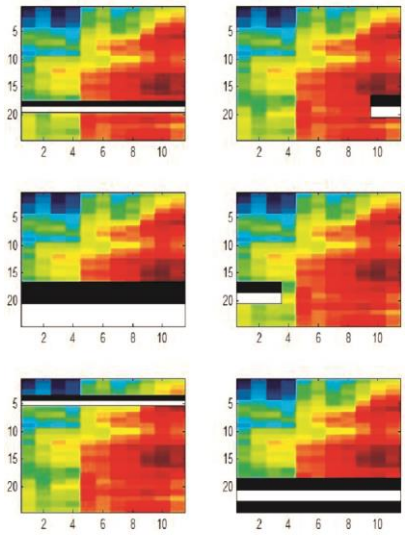
InIT Institut für angewandte
Informationstechnologie

Talkalyzer: Neue Algorithmen für die automatische Sprecher-Erkennung

Diplomand
Jan Stampfli

Dozierende
Mark Cieliebak
Thilo Stadelmann

Die heute gängigen Verfahren der automatischen Sprechererkennung befinden sich auf einem Stand, der den Einsatz in kommerziellen Bereichen ermöglicht. Werden aber die Bedingungen, beispielsweise durch fehlendes Wissen über die Anzahl und Identität vorhandener Sprecher, zu komplex, bietet der heutige Stand der Verfahren keine praktikable Verwendung. Dieser Arbeit vorangegangene Forschung zeigt, dass im zeitlichen Verlauf von Audiosignalen wichtige sprecherspezifische Information enthalten ist, die aber in den gängigen Systemen vernachlässigt wird.
In der vorliegenden Arbeit werden verschiedene Ansätze hinsichtlich der Integration zeitlicher Aspekte von Audiosignalen untersucht, mit dem Ziel, die bestehenden Verfahren der automatischen Sprechererkennung zu verbessern. Die Erkennungsleistung des aus diesen Ansätzen erarbeiteten Konzepts wird anhand von Clustering-Experimenten verifiziert. Den Experimenten liegen Audiodaten aus dem TIMIT-Datensatz zugrunde.
Das erarbeitete Konzept sieht vor, dass die zeitliche Information von Audiosignalen anhand eines kleinen Sets von 30 Filtern aus Spektrogrammen extrahiert wird. Zur Selektion dieses Filtersets wird auf eine Vorgehensweise aus der automatischen Objekterkennung (Computer Vision) zurückgegriffen. Grundsätzlich werden aus einem grossen Set von potentiellen Filtern mit Hilfe des AdaBoost-Algorithmus die entscheidenden Filter selektiert. Während des Clusterings werden zunächst die Audiodaten der Sprecher geladen und daraus Spektrogramme erstellt. Weiter werden mit dem selektierten Filterset die Sprechermerkmale aus den Spektrogrammen extrahiert und anhand dieser Merkmale probabilistische Sprechermodelle erzeugt. Abschliessend werden die Distanzen zwischen den Modellen berechnet und anhand dieser Distanzen die Modelle zu Clustern zusammengefasst. Die Ergebnisse wurden verglichen mit denen eines Baseline-Ansatzes, der die gängigsten Verfahren der automatischen Sprechererkennung beinhaltet. Der neue Ansatz erreicht für ein Clustering von 40 unterschiedlichen Sprechern eine Präzision von 87.37% und eine Ausbeute von 83.05%; der Baseline-Ansatz erzielte 89.10% Präzision und 87.24% Ausbeute. Obwohl die Ergebnisse zeigen, dass noch keine Verbesserung der bestehenden Verfahren erzielt werden konnte, bietet diese Arbeit dennoch eine Grundlage für weiterführende Forschung. Sie zeigt, welche Schritte des erarbeiteten Konzepts Anpassung benötigen, um die Erkennungsleistung entscheidend zu steigern.



Grafische Darstellung einer Auswahl von sechs Filtern aus dem berechneten Filterset, das zur Extraktion der Sprechermerkmale eingesetzt wird. Die Filter sind innerhalb von Spektrogrammen als schwarze und weisse Flächen dargestellt. Die horizontale Achse der Spektrogramme beschreibt den zeitlichen Signalverlauf, die vertikale Achse den Frequenzbereich, und "heissere" Farben die Lautstärke.