Ein Demonstrator zum Keyword-Spotting basierend auf gehörangepassten Audiomerkmalen

Dirk von Zeddelmann, Sebastian Urrigshardt

Abteilung Kommunikationssysteme
Fraunhofer FKIE
Fraunhoferstr. 20
53343 Wachtberg
dirk.von.zeddelmann@fkie.fraunhofer.de
sebastian.urrigshardt@fkie.fraunhofer.de

Abstract: In dieser Zusammenfassung stellen wir einen Demonstrator zur Schlüsselworterkennung (Keyword-Spotting) vor, der ohne vorheriges Training der zu detektierenden Schlüsselwörter auskommt. Die zugrundeliegende Technologie basiert auf der Verwendung von unüberwachten Matchingstrategien zusammen mit speziell für diese Anwendung entwickelten gehörangepassten Audiomerkmalen. Diese weisen eine größere Sprecherinvarianz auf als bisher bekannte Merkmale wie MFCCs. Die vorgestellte Technologie eignet sich besonders für Anwendungen, bei denen für das Keywordspotting nur unzureichendes Trainingsmaterial zur Verfügung steht und die zu detektierenden Schlüsselwörter eine gewisse Mindestlänge aufweisen.

1 Motivation

Klassische Ansätze des Keyword-Spottings basieren darauf, dass die verwendeten Erkennungsalgorithmen (etwa Hidden Markov Modelle oder Neuronale Netze) in einem Vorverarbeitungsschritt anhand von Referenzdaten auf die zu erkennenden Wörter trainiert werden. Der Trainingsschritt bei diesem *überwachten* Vorgehen ist einerseits zeitaufwendig und stellt andererseits signifikante Anforderungen an die Verfügbarkeit von Trainingsdaten. In zahlreichen Szenarien, wie beispielsweise dem akustischen (Langzeit-) Monitoring in realen Umgebungen, der Suche in Audiodatenbanken oder der Auswertung von sicherheitsrelevanten Überwachungsaufnahmen stehen jedoch nicht immer ausreichend Ressourcen zur Verfügung. Neben dem prinzipiellen Fehlen von geeignetem Trainingsmaterial für die zu erkennenden Worte ist ein weiteres Problem häufig das Fehlen geeigneter Annotationen oder, im Falle bestimmter Fremdsprachen, sogar die fehlende Möglichkeit diese Annotationen mit vertretbarem Aufwand zu erzeugen.

Vor diesem Hintergrund stellen wir einen unüberwachten Ansatz zur Detektion von Worten und Wortfolgen vor, welcher keinen Trainingsschritt benötigt. Hierzu berechnen wir aus den zu erkennenden Schlüsselworten zunächst sprecherunabhängige Merkmalsvektoren. Indem wir weiterhin voraussetzen, dass die Schlüsselworte eine gewisse Mindestlänge

aufweisen erreichen wir, dass Folgen solcher Merkmalsvektoren einen charakteristischen Fingerabdruck des Schlüsselwortes darstellen.

2 Unüberwachtes Keyword-Spotting

Bei dem hier vorgestellten Ansatz ist es von grundlegender Bedeutung, die zu verarbeitenden Sprachsignale in Folgen von Merkmalsvektoren zu überführen, deren zeitliches Verhalten mit den enthaltenen Lautfolgen korreliert. Experimente zeigen, dass gebräuchliche Sprachmerkmale wie MFCCs hierfür nicht optimal geeignet sind. Skrowonski und Harris [SH04] verwenden darum im Schritt der Merkmalsextraktion in einer Analysefilterbank spezielle, bezüglich der kritischen Bandbreiten der menschlichen Hörwahrnehmung angepasste Bandbreiten. Weiterhin zeigen Untersuchungen, dass gebräuchliche Parametersätze bezüglich der Zeitauflösung nicht gut auf die Detektion von Lautfolgen angepasst sind. Zur besseren Anpassung schlugen wir in einer Vorgängerarbeit [vZKM10] einen aus der Musikerkennung abgeleiteten Ansatz unter Verwendung von Kurzzeitstatistiken vor.

Innerhalb des zum eigentlichen Keyword-Spotting verwendeten Matching-Verfahrens wird die aus einem Schlüsselwort gewonnene Merkmalsfolge mit der zu durchsuchenden Datenbank korreliert. Hierbei entsteht eine Score-Funktion, die den Grad der Übereinstimmung des Schlüsselwortes mit jeder Position der Merkmalsdatenbank angibt. Innerhalb der Score-Funktion äußern sich Übereinstimmungen in lokalen Maxima. Zur Kompensation variabler Wortdauern wird zur Bestimmung der Score-Funktion eine modifizierte Form des Teilfolgen-DTWs (Dynamic Time Warping) verwendet [vZKM10].

3 Demonstrator

Abbildung 1 zeigt die Programmoberfläche des am FKIE entwickelten Keyword-Spotting Demonstrators. Innerhalb des Demonstrators lassen sich zu durchsuchende Audioaufzeichnungen einladen und mittels eines integrierten Audioplayers wiedergeben (Abbildung 1 oben). Im linken Abschnitt der Oberfläche sind die zu suchenden Schlüsselwörter als Zeitsignal und als Merkmalsfolge abgebildet. Neue Suchwörter lassen sich interaktiv in das Programm laden. Nach Starten des Analysevorgangs werden die vom Programm detektierten Zeitabschnitte auf der rechten Seite dargestellt. Die Ergebnisse werden im gegebenen Beispiel ihrer Ähnlichkeit zu der Suchanfrage entsprechend geordnet. Eine Wiedergabefunktion erlaubt das Abspielen der detektierten Ergebnisse, wobei die Position innerhalb der Aufzeichnung, an der ein Schlüsselwort erkannt wurde, farblich markiert wird. Zusätzlich besteht die Möglichkeit, innerhalb des Detektionsprozesses den Score-Wert zu bestimmen, um beispielsweise nur Treffer einer bestimmten prozentualen Ähnlichkeit zu der Anfrage anzeigen zu lassen.

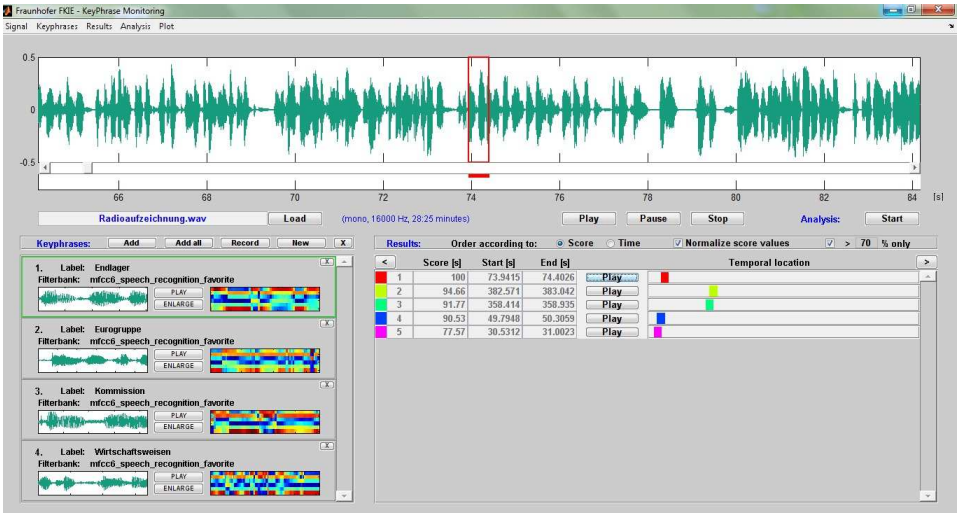

Abbildung 1: FKIE Keyword-Spotting Demonstrator

Literatur

[SH04] Mark D. Skowronski und John G. Harris. Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition. *The Journal of the Acoustical Society of America (JASA)*, 116(3):1774–1780, 2004.

[vZKM10] Dirk von Zeddelmann, Frank Kurth und Meinard Müller. Perceptual Audio Features for Unsupervised Key-Phrase Detection. In *Proc. IEEE ICASSP*, Dallas, TX, USA, Marz 2010.