Untersuchung eines Algorithmus zur Eigenspracherkennung über Signal-Kohärenz in Labor- und Alltagssituationen

Sascha Bilert, Jörg Bitzer, Inga Holube

Institut für Hörtechnik und Audiologie, Jade Hochschule Oldenburg

Schlüsselwörter: Own-Voice-Detection, Kohärenz, Ecological Momentary Assessment, DiapixUK, Smartphone-Messsystem

Einleitung

Um den akustischen Alltag eines Probanden über einen längeren Zeitraum gut beschreiben zu können, sollten neben subjektiven Bewertungen von wahrgenommenen Eindrücken auch objektive Messdaten, wie z.B. Pegel und Spektren, in alltagsnahen Situationen erfasst werden. Bei der Aufzeichnung dieser objektiven Daten müssen unterschiedliche Herausforderungen betrachtet werden. Hierzu gehört zum einen die Berücksichtigung von § 201 StGB (Verletzung der Vertraulichkeit des Wortes) und zum anderen die Identifikation der Zeiträume, in denen die Probanden selbst sprechen, also nicht allein die akustische Umgebung erfasst wird. Der nachfolgende Beitrag fokussiert sich auf die Evaluation eines Algorithmus zur Erkennung der Eigensprache (own-voice-detection, OVD) über die binaurale Signal-Kohärenz. Der betrachtete OVD-Algorithmus ist hierbei ein erster Ansatz, um Eigensprachsegmente anhand von privatsphären-erhaltenden extrahierten Merkmalen zu identifizieren. Für die Untersuchungen wurden Messreihen mit einem Kunstkopf und mit Probanden unter Laborbedingungen betrachtet.

Messsystem

Bitzer et al. (2016) stellten ein Messsystem zur Erfassung von Alltagssituationen vor, das die Vertraulichkeit des gesprochenen Wortes gewährleistet. Dieses Messsystem wurde zur Verbesserung der Nutzbarkeit weiterentwickelt. Die aktualisierte Version besteht aus zwei Mikrofonen, einer Bluetooth-Übertragungseinheit und einem Smartphone (Kowalk et al., 2018). Die MEMS-Mikrofone (InvenSense, Typ INMP504) wurden jeweils am rechten und linken Brillenbügel fixiert. Sie waren über ein Kabel mit der Bluetooth-Übertragungseinheit verbunden. In dieser Einheit wurden die Mikrofonsignale mit einer Samplingrate von f_S = 48 kHz und einer Quantisierung von 16 Bit analog-digital-gewandelt. Die Audiodaten wurden anschließend via Bluetooth (A2DP) an das Smartphone übermittelt und dort weiterverarbeitet. Um die Vertraulichkeit des Wortes zu gewährleisten, wurden aus den Audiodaten Merkmale (Feature) extrahiert und die aufgezeichneten Audiodaten anschließend gelöscht. Zu den Featuredaten gehörten der Effektivwert, die Nulldurchgangsrate sowie das Auto- und Kreuzleistungsdichtespektrum (A- bzw. KLDS). Die Speicherung der vollständigen LDS würde eine Rekonstruktion der Sprache ermöglichen. Deshalb wurden gemittelte LDS (Mittelungszeit 125 ms, bei einem Blockvorschub von 12,5 ms und einer Blocklänge von 25 ms) nur alle 125 ms gespeichert.

Parallel zur Aufzeichnung der objektiven Daten, können in einer Langzeitanwendung subjektive Probandenbewertungen über ein Ecological Momentary Assessment (EMA) gesammelt werden (Stone und Shiffman, 1994). Die EMA kann hierbei zur Nutzerfreundlichkeit direkt auf dem Smartphone-Messsystem durchgeführt werden, so dass eine parallelisierte Aufzeichnung der subjektiven wie auch der objektiven Daten und deren anschließender Auswertung (offline) erfolgt.

Methoden

Im Folgenden wird auf den verwendeten OVD-Algorithmus, den Entwurf eines adaptiven Schwellenkriteriums und die unterschiedlichen Messumgebungen, welche zur Evaluation des OVD-Algorithmus verwendet wurden, eingegangen.

OVD-Algorithmus:

Das Ziel des OVD-Algorithmus ist die Identifikation von eigensprachbehafteten Featuredaten, um beispielsweise eine genauere Schätzung des Hintergrundpegels in Alltagssituationen zu gewährleisten. Dazu wurde der OVD-Algorithmus von Bitzer und Kissner (2016) evaluiert. Dieser Algorithmus arbeitete auf Basis der spektralen A/KLDS. Für die Identifikation der Eigensprachanteile wurde die Kohärenz zwischen dem linken und rechten Kanal mit den beschriebenen gesampelten Featuredaten geschätzt. Anschließend erfolgte eine Mittelwertbildung des Realteils der geschätzten Kohärenz im Frequenzbereich von 400 Hz bis 1000 Hz. In Abbildung 1 ist ein Ausschnitt einer Unterhaltung zwischen zwei Probanden in einer ruhigen Umgebung

dargestellt. Die obere Grafik zeigt den frequenzabhängigen Realteil der Kohärenz über der Zeit, während die untere Abbildung den gemittelten Einzahlwert der Kohärenz widerspiegelt (blau). Die rote Kurve gibt den exponentiell geglätteten Verlauf der gemittelten Kohärenz mit $\tau = 1{,}0$ s an. Zur anschließenden Trennung zwischen Eigensprachsegmenten und Fremdgeräuschen wurde bisher ein konstantes Schwellenkriterium von $\sigma_{fest} = 0{,}6$ verwendet.

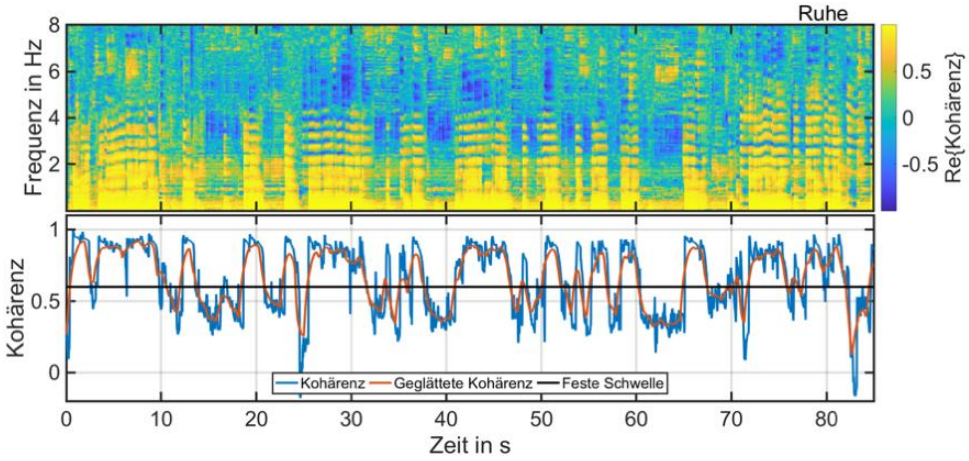

Abbildung 1: Exemplarische Darstellung des Realteils eines geschätzten Kohärenzspektrums (oben) und der gemittelten Einzahlwerte der Kohärenz im Frequenzbereich zwischen 400 und 1000 Hz (unten) über einem Zeitbereich von 85,0 s. Der Einzahlwert ist sowohl ungeglättet (blau) als auch exponentiell geglättet (rot) abgebildet. Die schwarze Linie (unten) kennzeichnet das bisher verwendete konstante Schwellenkriterium.

Konstantes vs. adaptives Schwellenkriterium:

Pilotuntersuchungen im Labor bestätigten die Verwendung des konstanten Schwellenkriteriums zur Trennung zwischen Eigensprach- und Fremdsprachanteilen in ruhigen Umgebungen, ließen jedoch Grenzen in der Anwendbarkeit bei Variation des Hintergrundpegels erwarten. Deshalb wurde als alternativer Ansatz zu dem konstanten Schwellenkriterium ein adaptives Schwellenkriterium verwendet.

In Abbildung 2 ist ein Zeitausschnitt der ungeglätteten und der geglätteten mittleren Kohärenz bei einem Hintergrundpegel von 70 dB(A) dargestellt. In der gleichen Grafik sind auch das feste und das adaptive Schwellenkriterium abgebildet. Zur Berechnung des adaptiven Schwellenkriteriums wurde das gleitende Minimum und das gleitende Maximum in einem Zeitfenster von 25 s bestimmt und der Mittelwert aus diesen beiden Werten gebildet. Der Mittelwert wurde anschließend noch um 25 % des Abstands zwischen Minimum und Maximum angehoben. Mit diesem Berechnungsverfahren ergab sich ein Schwellenkriterium, welches sich zeitlich an die Veränderung des Schwankungsbereichs der gemittelten Kohärenz anpasst.

Labor- und Alltagssituationen:

Zur Evaluation der beiden Schwellenkriterien wurden zwei Messreihen durchgeführt. In der ersten Messreihe wurden 51 unterschiedliche Messkonfigurationen unter kontrollierten Messbedingungen mit Hilfe eines Kunstkopfes mit Mundsimulator (KEMAR, G.R.A.S Head and Torso Type 45BM) und dem Smartphone-Messsystem im TASCARpro-Aufbau (Grimm et al., 2015) aufgenommen. Der KEMAR befand sich immer im Zentrum des TASCARpro-Aufbaus. Für das Testsignal wurde eine Unterhaltung zwischen dem KEMAR und einem Lautsprecher des TASCARpro-Systems mit der männlichen und der weiblichen Fassung des Oldenburger Satztests (OLSA, Wagener et al., 1999, Wagener et al., 2014) simuliert. Als Störsignal wurde ein Cafeteria-Szenario über das TASCARpro-System wiedergegeben. In den Messkonfigurationen wurden unter anderem unterschiedliche Signal-Rausch-Abstände, Sender-Empfänger-Distanzen, Sprachpegel und Azimut- und Elevationswinkel untersucht.

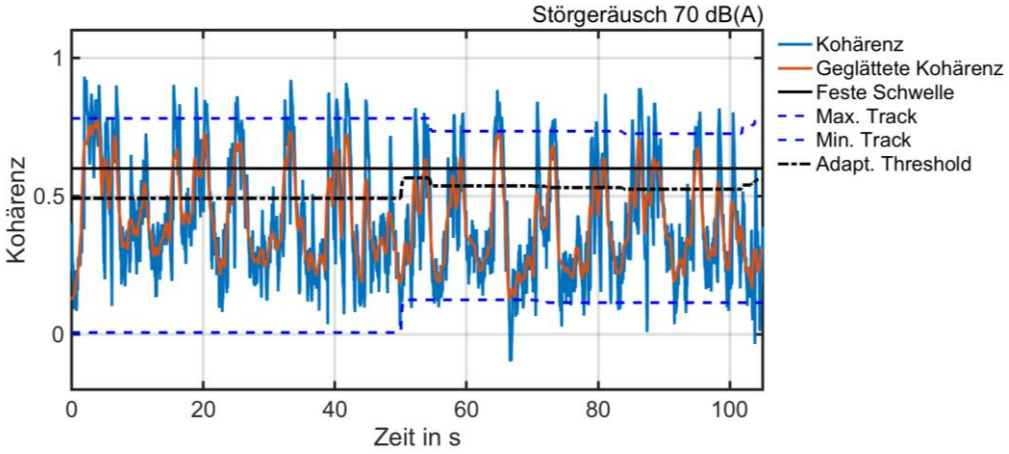

Abbildung 2: Gemittelte Schätzung der Kohärenz in der geglätteten und ungeglätteten Version bei einem Hintergrundpegel von 70 dB(A). Die schwarze durchgezogene Linie kennzeichnet das feste Schwellenkriterium, während die schwarze gestrichelte Linie das adaptive Kriterium markiert.

Die zweite Messreihe beinhaltete eine Probandenstudie mit $N = 8$ Teilnehmern, wovon jeweils die Hälfte männlich bzw. weiblich war. Ein Proband, der das Smartphone-Messsystem trug, befand sich immer im Zentrum des TASCARpro-Aufbaus, während ein anderer Proband in einem festgelegten Abstand saß oder stand. Bei unterschiedlichen Hintergrundgeräuschpegeln und verschiedenen Abständen führten die Probanden eine Unterhaltung durch. Die Konversation zwischen den Probanden wurde mit Hilfe der DiapixUK-Bilder (Baker und Hazan, 2011) oder über vorgegebene Themen angeregt und betrug zwischen 2 und 5 min. Die Aufgabe bei der Verwendung der DiapixUK-Bilder bestand daran, zwölf Unterschiede in jeweils einem Bildpaar zu finden.

Ergebnisse

Für die zwei Messreihen wurden die Erkennung der Sprache derjenigen Person, die das Smartphone-System trug, bzw. des KEMAR bei Verwendung des festen und des adaptiven Schwellenkriterium analysiert. In Abbildung 3 ist für die erste Messreihe die Verteilung der ungeglätteten KEMAR- und GEGENSPRECHER- (Lautsprecher)-Kohärenz bei unterschiedlichen Messkonfigurationen dargestellt. In der künstlichen Laborumgebung ändert sich je nach Messkonfiguration der Schwankungsbereich der Kohärenz, weshalb mit der adaptiven Schwelle, im Vergleich zur konstanten Schwelle, eine bessere Trennung zwischen den beiden Sprachquellen erreicht werden kann.

Die zweite Messreihe zeigt die Erkennungsrate der Eigensprache (Own-Voice Hit-Rate) unter Betrachtung unterschiedlicher Hintergrundgeräuschpegel während der Unterhaltung zwischen zwei Probanden. In Abbildung 4 ist die Hit-Rate für die ungeglättete (blau) und für die geglättete (rot) gemittelte Kohärenz dargestellt. Die Erkennungsrate steigt für beide Kohärenzwerte durch die Verwendung des adaptiven Schwellenkriteriums im Vergleich zum konstanten Schwellenkriterium an, während die Hit-Rate bei dem festen Schwellenkriterium durch den steigenden Hintergrundpegel abnimmt. Außerdem ist eine Verringerung der Standardabweichung, die die Streuung zwischen den Probanden beschreibt, unter Verwendung der adaptiven Schwelle zu sehen.

Zusammenfassung

Zur Erkennung der Zeitabschnitte, in denen der Nutzer des Smartphone-Systems zur Alltagserfassung selbst spricht, wurde ein Schwellkriterium vorgeschlagen, das auf dem gemittelten Realteil der Kohärenz beruht. Mit zwei verschiedenen Messreihen konnte gezeigt werden, dass eine Erkennung der Eigensprachsegmente über Featuredaten generell möglich ist und dass die Erkennungsrate durch den Einsatz eines adaptiven Schwellenkriteriums bei beiden Messreihen gegenüber einem konstanten Schwellenkriterium gesteigert werden kann.

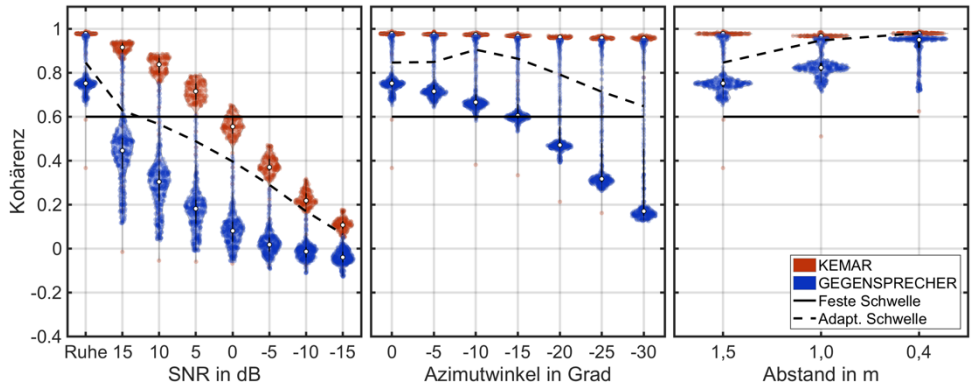

Abbildung 3: Verteilung der KEMAR- und GEGENSPRECHER-Kohärenz in drei unterschiedlichen Messkonfigurationen mit unterschiedlichen Einstellungen. Das konstante (durchgezogen) und das adaptive (gestrichelt) Schwellenkriterium zeigen die Trennbarkeit in den jeweiligen Messkonfigurationen.

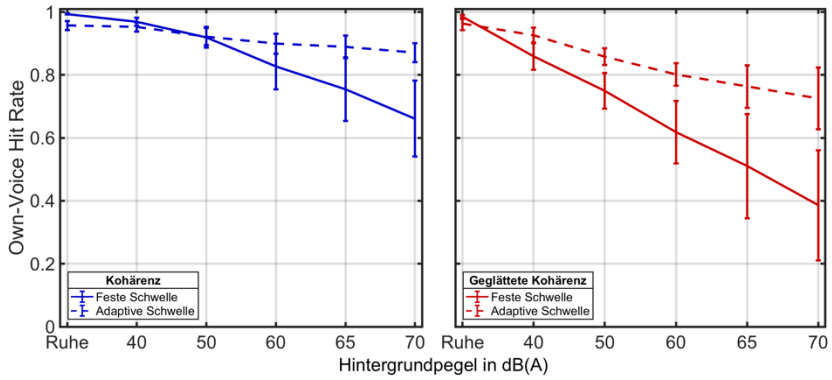

Abbildung 4: Own-Voice Hit Rate für die ungeglättete (blau) und die geglättete (rot) mittelte Kohärenz. Die Hit Rate wurde sowohl mit dem festen, wie auch mit dem adaptiven Schwellenkriterium berechnet.

Danksagung

Die Untersuchungen wurde durch das Niedersächsische Vorab der VW-Stiftung im Rahmen des Forschungsschwerpunktes „**H**ören im **All**tag **O**ldenburg“ (HALLO) und das Hearing Industry Research Consortium (IRC, Projekt IHAB-RL) gefördert.

Literatur

Baker R & Hazan V (2011) DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs, Behavior Research Methods 43(3), 761–770

Bitzer J, Kissner S, Holube I (2016) Privacy-aware acoustic assessment of everyday life, Journal of the Audio Engineering Society 64(6), 395-404

Bitzer J & Kissner S (2016) Two-channel coherence-based own voice detection for privacy-aware long-term acoustic measurements, ITG-Fachbericht 267: Speech Communication, 170-174

Grimm G, Luberadzka J, Herzke T, Hohmann V (2015) Toolbox for acoustic scene creation and rendering (TASCAR): Render methods and research applications, Proceedings of the Linux Audio Conference.

Kowalk U, Kissner S, von Gablenz P, Holube I, Bitzer J (2018) An improved privacy-aware system for objective and subjective ecological momentary assessment, Proceedings of the International Symposium on Auditory and Audiological Research, Nyborg, Dänemark

Stone A A & Shiffman S (1994) Ecological momentary assessment (EMA) in behavorial medicine, Annals of Behavioral Medicine 16, 199-202

Wagener K, Brand T, Kollmeier B (1999) Entwicklung und Evaluation eines Satztests für die deutsche Sprache – Teil III: Evaluation des Oldenburger Satztests, Zeitschrift für Audiologie 38(3), 86-95

Wagener K, Hochmuth S, Ahrlich M, Zokoll M, Kollmeier B (2014) Der weibliche Oldneburger Satztest. 17. Jahrestagung der Deutschen Gesellschaft für Audiologie