R-De-03

Erfassung von Schläfrigkeit in Funkverkehr-gestützter Kommunikation: Sprachsignal-basierte Mustererkennungsanalyse phonetisch-linguistischer Merkmale

Sebastian Schnieder[1], Elmar Nöth[2], Anton Batliner[2], Sonja-Dana Roelen[1], Anke Brunstein[3] und Jarek Krajewski[4]

[1] Bergische Universität Wuppertal
[2] Friedrich-Alexander-Universität, Erlangen-Nürnberg
[3] Technische Universität Berlin
[4] Bergische Universität Wuppertal/Rheinische FH Köln

Schlüsselwörter: Schläfrigkeit, Sprache, Funkverkehr, Schläfrigkeitsdetektion

Zusammenfassung

Die Detektion von kritischen Schläfrigkeitszuständen stellt aus der Perspektive von Unfallprävention, Komfortsteigerung und Optimierung von Arbeitsleistung sowohl im betrieblichen als auch privaten Umfeld eine wertvolle Bereicherung dar. Ziel der Untersuchung ist die Identifikation schläfrigkeitsinduzierter phonetisch-linguistischer Veränderungen der Sprache und die Entwicklung eines darauf aufbauenden, automatisierten Schläfrigkeitsmessverfahrens. Zu diesem Zweck wird ein schlafdeprivationsbasiertes Sprachkorpus (N = 89) aufgezeichnet. Aufbauend auf Fortschritten der mustererkennungsbasierten Sprachemotionserkennung werden sowohl ein hybrides brute-force wie auch ein theoriegeleitetes Merkmalset extrahiert. Die Kernergebnisse der Untersuchung sind: Aufbau eines realitätsnahen, moderate Schläfrigkeitsintensitäten berücksichtigenden Korpus, zusätzliche Nutzung von Video- statt unimodalen Audio-Annotationen, Korrektur der bisher erzielten Erkennungsraten von Schläfrigkeit aus Sprache und bessere Detektionsraten für männliche Probanden. Darüber hinaus konnte die Bedeutung von Expertise und Kontextinformation bei der Annotation von Schläfrigkeit durch Naive Rater, Experten und Versuchsleitern festgehalten werden. Des Weiteren wurden eine Vielzahl von Artikulations- und Stimmqualitätsbezogenen akustischen Korrelaten von Schläfrigkeit identifiziert.

Zentrale Einsatzfelder sprachlicher Schläfrigkeitsdetektion liegen in kommunikationsintensiven Tätigkeiten (z.B. bei Fluglotsen oder bei Funkverkehr gestützten Arbeitsplätzen im Allgemeinen) bei der auf eine bereits vorhandene Kommunikationsinfrastruktur zurückgegriffen werden kann. Ferner ist der Einsatz auch in sprachgesteuerten Human-Computer-Interaction z.B. im Rahmen von Fahrerassistenzsystemen oder Assisted Living Anwendungen denkbar, um empathischere Dialogführung zu ermöglichen.

Das Ziel der Untersuchung ist die Identifikation schläfrigkeitsinduzierter phonetisch-linguistischer Veränderungen der Sprache und die Entwicklung eines darauf aufbauenden, automatisierten Schläfrigkeitsmessverfahrens. Zum Aufbau dieses Systems wird ein schlafdeprivationsbasiertes Sprachkorpus erstellt. Die Ausgangsfragen des Forschungsprojektes beziehen sich somit auf eine hybride ingenieurwissenschaftliche datengeleitete „brute-force“ und eine phonetische wissensbasierte Perspektive. Beide Perspektiven des „Getting Better“, also der Optimierung von Schläfrigkeitsklassifikationsraten, und des „Getting Wiser“, der theoriegeleiteten Identifikation von phonetischen Einzelmerkmalen der Stimme, sollen daher im Rahmen der Untersuchung aufgegriffen werden.

Die Zielsetzungen lassen sich summierend wie folgt beschreiben:

- Aufbau eines realitätsnahen Schläfrigkeitskorpus mit alltagsrelevanter Verteilung von Schläfrigkeitszuständen, Korrektur der bisher erzielten Erkennungsraten von Schläfrigkeit aus Sprache.
- Anwenden von State-of-the-Art Merkmalsextraktions- und Mustererkennungsverfahrens auf Schläfrigkeitsdaten zur Bestimmung von realistischen Detektionsraten,
- Verifikation von in bisherigen Untersuchungen identifizierten akustischen Schläfrigkeitskorrelaten,
- Überprüfung der in der Paralinguistik häufig genutzten Audio-Annotationen gegenüber Videoannotationen oder Audio-Video-Annotationen,
- Anwendbarkeit von naiven Ratern vs. geschulten Experten zur Schläfrigkeitsannotation,
- Spezifizierung von männlichen und weiblichen Submodellen der akustischen Schläfrigkeitsmessung. Hinsichtlich der Geschlechterunterschiede in der Detektion von Schläfrigkeit ist zu erwarten, dass sich die Schläfrigkeit von männlichen Probanden besser detektieren lässt, da Frauen ihre Schläfrigkeit in der Stimme weniger zeigen und eher zu kanonischer Aussprache neigen (Trudgill, 1972; Kreiman & Sidtis, 2011).

Methode

Zum Aufbau der Sprachkorpora wurden Studenten der Bergischen Universität Wuppertal als Probanden akquiriert. Zusätzlich wurden die sich in ihrer Stimmwirkung mit Schläfrigkeit überlappenden Depressionszustände ermittelt. Ferner wurden eine Reihe von Drittvariablen erfasst, die neben der Schläfrigkeit Einfluss auf die Stimmcharakteristik nehmen wie Geschlecht, Alter, Größe, Gewicht, Raucherstatus, Erkältungsstatus und regionaler Dialekt. Die Stichprobengröße lag bei N = 89 (23 m, 66 w) Probanden (Alter 23.7 Jahre ± 5.2).

Schläfrigkeit wurde bei den Probanden durch partielle Schlafdeprivation mit einem verkürzten Nachtschlaf (2 Stunden Schlaf zwischen 00:00 bis 02:00 Uhr) am Vortag der Messungen induziert und durch Aktimetrie überprüft. Zusätzlich wurden die Probanden aufgefordert, von 02:00 Uhr bis zum Beginn der Messungen halbstündig E-Mails zu beantworten. Messungen der Probanden erfolgten maximal drei Mal am Tag, zu jeweils 09:30 Uhr, 13:30 Uhr und 17:30 Uhr. Die Dauer für einen Probanden lag bei ungefähr 2,5 Stunden und wurde als Laboruntersuchung durchgeführt. Versuchsleiter und Proband waren während des gesamten Versuches lediglich über ein Headset miteinander verbunden, um die Deaktivierung einer sozialen Isolationssituation zu nutzen (s. Abbildung 1). Nach dem Ausfüllen einer Fragebogenbatterie wurden dem Proband ein EKG-Brustgurt angelegt und das Eyetrackingsystem kalibriert (s. Abbildung 2). Anschließend folgten ca. eineinhalbstündige Sprechaufgaben, die im weiteren Abschnitt beschrieben werden.

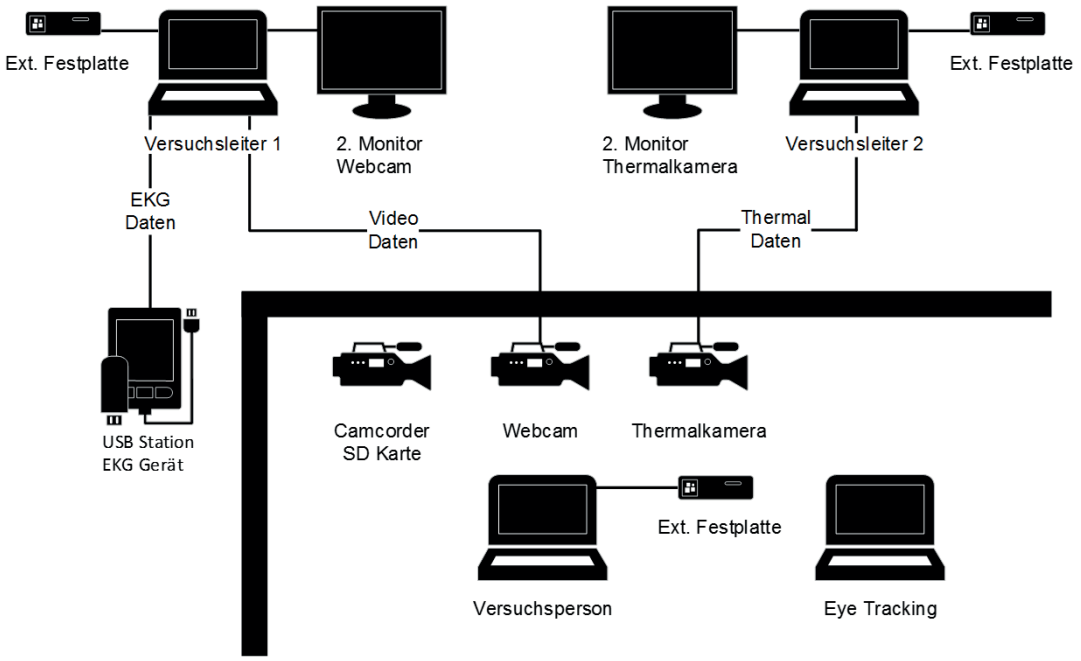

Abb. 1: Technische Aufnahme-Infrastruktur des Probanden-Messplatzes und des Versuchsleiter-Kontrollraums.

Sprechaufgaben

Während der Messung unter Laborbedingungen wurden Anwendungssituationen in spezifischen Kontexten via Bildschirmpräsentation und durch mündliche Anweisungen des Versuchsleiters simuliert, wie z.B. Fluglotsenaufgaben mit typischen Sprechsituationen („T-C-seven, turn left heading one-eight-zero degrees"), Aufgaben aus der Fahrerassistenz („Ich suche die Friesenstraße"), Leseaufgaben auf Deutsch und Englisch (Zeitungsartikel, Texte aus dem Roman „Homo Faber", phonetisch ausbalancierte Texte wie „Der Nordwind und die Sonne") sowie Dialogaufgaben mit Kommandosprache, wie sie im Kontext von Assisted-Living Systemen vorkommen („Computer, please make espresso").

Abb. 2: Versuchssetting mit Probandenmessplatz. Aufgebaut sind die Kamera- (Infrarotkamera, HD-Videokamera), Sprachrecording-, Eyetracking- und Fahrsimulationssysteme.

Instrumente

Als Schläfrigkeitsreferenzwert dient der Karolinska Sleepiness Scale (KSS) Schläfrigkeitsscore. Der Selbstbericht wurde direkt im Anschluss an die jeweiligen Sprechaufgaben erhoben. Die Fremdberichte fanden im Anschluss an die Messung in Form einer auf audio-visuellen Daten basierenden KSS-Einschätzung statt. Das jeweilige audio-visuelle Probandenmaterial wurde in Sprechaufgaben gesplittet und dem Rater in randomisierter Reihenfolge präsentiert, Informationen über Uhrzeit der Messung, Schlafdeprivation und Selbstberichte der Probanden waren den Beurteilern nicht bekannt. Vier (2 m, 2 w) unabhängige, wenig geschulte Beurteiler bewerteten jeweils das vollständige Material des ersten Probandenblockes hinsichtlich ihrer KSS-Werte. Weitere vier (2 m, 2 w) unabhängige Beurteiler bewerteten in identischem Prozedere den zweiten Probandenblock der restlichen Teilnehmer.

Ergebnisse

Univariate Korrelate

Die bisherigen Ergebnisse zur akustischen Schläfrigkeitsdetektion basieren auf realitätsfernen bimodalen Extremverteilungen und alltagsfernen Extremintensitäten der Schläfrigkeit (Schuller et al., 2011). Das hier erstellte Korpus liefert auf Grundlage eines umfangreichen Probandenkollektivs hingegen eine realistische Einschätzung der akustischen Erkennung von Schläfrigkeit (s. Abbildung 3). Dieses hohe Maß an Realismus drückt sich in eher niedrigeren Mittelwerten und Streuungen der Schläfrigkeitswerte aus. Eine Konsequenz dieser realistischen Datengrundlage ist eine Korrektur der bisher erzielten Erkennungsraten von Schläfrigkeit aus Sprache. Aus der reduzierten Standardabweichung des im Projekt erstellten Korpus lässt sich über die Anwendung der Pearson Korrelationsformel die daraus resultierende Reduktion der Validitätskorrelation schätzen (0.5 * 1.4/2.4 = 0.3). Sie entspricht der vermuteten Reduktion von r = .5 für das extremverteilten SLC-Korpus (Schuller, 2011) auf den realistischen Wert von r = .3,

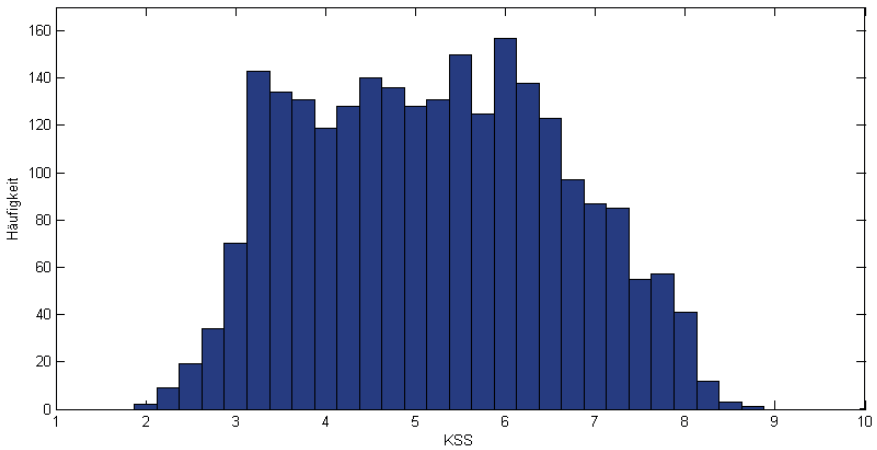

Abb. 3: Häufigkeitsverteilung der Audio-Video Rating basierten KSS Schläfrigkeitsratings. Die realitätsnahe Verteilung der Schläfrigkeit drückt sich in der Dominanz der mittleren Schläfrigkeitsklassen KSS 3 bis KSS 6 gegenüber den Extremereignissen KSS < 3 und KSS > 7 aus.

Eine geschlechtsspezifische Modellierung verbessert die Detektion von Schläfrigkeit geringfügig. Zunächst wurde auf den männlichen Aufnahmen eine höhere Validität (r = .54) beobachtet als auf den weiblichen (r =.44). Sogar wenn das Vorhersagesystem nur mit weiblichen Aufnahmen trainiert wurde, ergab sich dieses Bild (r = .51 vs. .44). In detaillierten Untersuchungen konnte gezeigt werden, dass dies nicht etwa darauf basiert, dass die akustischen Parameter für weibliche Sprache nicht ähnlich geeignet wären, sondern tatsächlich darauf, dass Frauen die Schläfrigkeit in ihrer Stimme weniger zeigen als Männer.

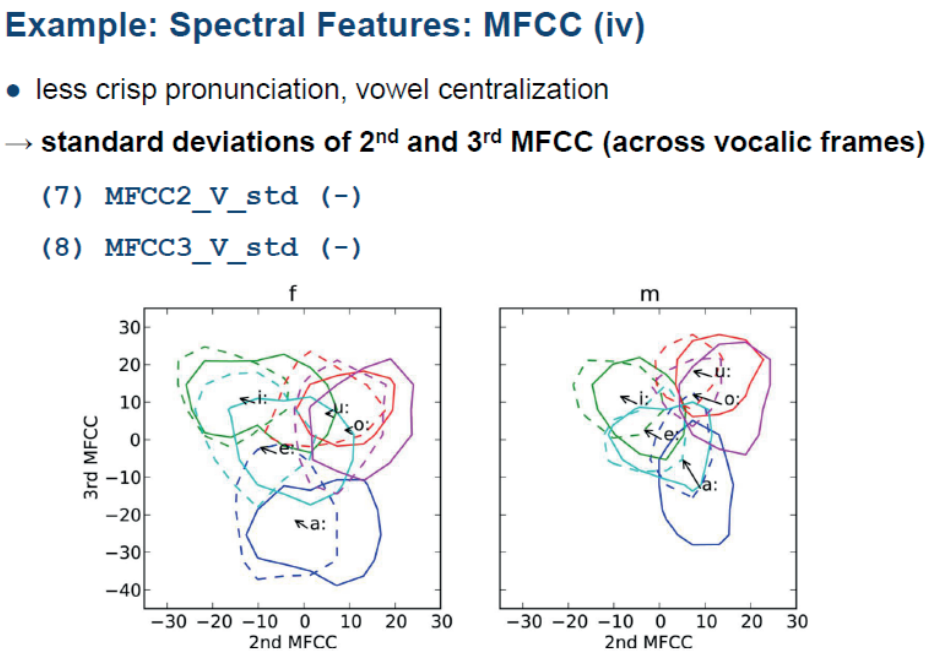

Abb. 4: Geschlechtsspezifische Verteilung von MFCC2 und MFCC3 Kennzahlen und ihre Veränderung bei eher niedriger und eher hoher Schläfrigkeit (strichliert). Sichtbar werden bei Schläfrigkeit Vokalzentralisation und eine reduzierte SD der Merkmale MFCC2 und MFCC 3, was einer reduzierten Artikulationspräzision entspricht; vgl. (Hönig 2014c).

Zahlreiche Veränderungen von artikulationsbezogenen Merkmale konnten festgestellt werden; Abbildung 4 illustriert diese Veränderung anhand der MFCC2 und MFCC3 Parameter. Aus 3705 akustischen Merkmalen wurden 27 Parameter theoriegeleitet ausgewählt, die die Bereiche Prosodie, Rhythmik und Stimmqualität abbilden. Diesen wird eine gleiche Anzahl an datengeleiteten (stepwise-forward-selection aus einem Benchmark-Feature-Set) Merkmalen gegenübergestellt (vgl. Hönig et. al., 2014). Wie zu erwarten war, erreicht die theoriegeleitete Auswahl nicht die Validität des „brute-force-Ansatzes", jedoch ist der Unterschied nicht zu groß (r = .33 vs. .41). Eine datengeleitete Auswahl von 27 aus allen 8073 Merkmalen führt ebenfalls zu r = .33. Hinsichtlich der erwarteten Veränderung der Korrelate mit erhöhter Schläfrigkeit ergab sich in den meisten Fällen ein stimmiges Bild, was als Bestätigung des Ansatzes zu werten ist. Bei Männern zeigt sich Schläfrigkeit eher in spektralen Merkmalen (Zentralisierung, d.h. verwaschenere Aussprache), bei Frauen eher in prosodischen Merkmalen (z.B. geringere Tonhöhe),

Tab. 1: Die wichtigsten theoriegeleiteten akustischen Korrelate und ihre Spearman-Korrelation rho zu Schläfrigkeit (KSS), jeweils für alle Sprecher und getrennt für weiblich (f) und männlich (m). Je stärker die Korrelation, desto dunkler ist die Zelle eingefärbt.

Subgroup	Feature	Sleepiness		
		all	f	m
Formants	(3) product of the standard deviations of F1 and F2	+0.02	−0.04	−0.14
MFCC	(5) average of second MFCC across vocalic frames	−0.24	−0.10	−0.44
	(7) std. deviation of second MFCC across vocalic segments	−0.04	+0.02	−0.23
	(8) std. deviation of third MFCC across vocalic segments	−0.18	−0.18	−0.29
	(10) std. deviation of third MFCC across consonantal seg.	−0.13	−0.09	−0.26
Pitch	(11) average of pitch estimates across vocalic frames	+0.04	−0.26	−0.11
	(13) standard deviation of syllables' average F0	−0.04	−0.07	+0.05
Duration	(20) average syllables' relative durations	+0.24	+0.22	+0.23
Voice Quality	(31) raw shimmer (local change of frame-to-frame loudness)	−0.33	−0.32	−0.26
	(33) spectral harmonicity	+0.01	−0.13	−0.15
	(34) spectral tilt, averaged over vocalic frames	+0.07	+0.09	+0.42

Die wichtigsten der manuell ausgewählten Korrelate sind in Tabelle 1 aufgeführt. Zur Interpretation der einzelnen Merkmale sei auf (Hönig 2014b) verwiesen,

Maschinelles Lernen

Die Anwendung von Support Vector Regression (SVR) ergibt auf Grundlage von 1759 Sprachdateien aus 37 unterschiedlichen Sprechaufgaben mit vier Audio-Video-Annotatoren auf ungesehenen Sprechern r = .20 (Frauen r = .18, Männer r = .26). In der sprecherabhängigen Variante erzielen wir r = .47 (Frauen r = .47, Männer r = .48). Nutzt man eine SVR auf Grundlage von 206 Sprachdateien aus zwei unterschiedlichen Sprechaufgaben mit zwei Audio-Annotationen (Trainings- und Testset mit Audio-Label) auf gesehenen Sprecher, konnten wir eine Validitätskorrelation von r = .20 (Frauen r = .14, Männer r = .33) erzielen. Für die gleiche Konfiguration erreichte wir für Video-Annotation eine verbesserte Validitätskorrelation von r = .29 (Frauen r = .27, Männer r = .37). In allen Analyse-Varianten wird deutlich, dass Männer besser zu detektieren sind als Frauen und sprecherabhängig gute Validitätskorrelationen erreicht werden. Zusätzlich ist die Video-Annotation der Audio-Annotation in Bezug auf die Validität überlegen.

Im Vergleich von unimodalen Video- gegenüber multimodalen Audio-Video Annotation zeichnet sich ein heterogenes Bild ab. Wie die Vergleiche von Interrater-Reliabilitäten gezeigt haben, ist die Expertise und Kontextinformation für eine hohe Reliabilität der Schläfrigkeitsratings entscheidend. Entsprechend steigert diese hohe Qualität von Annotationen von Versuchsleitern mit hoher Expertise und Kontextwissen zusätzlich die Validität (vgl. hohe Interrater-Reliabilität und Validität der Interspeech 2011 Speaker State Challenge) (Schuller et al., 2011).

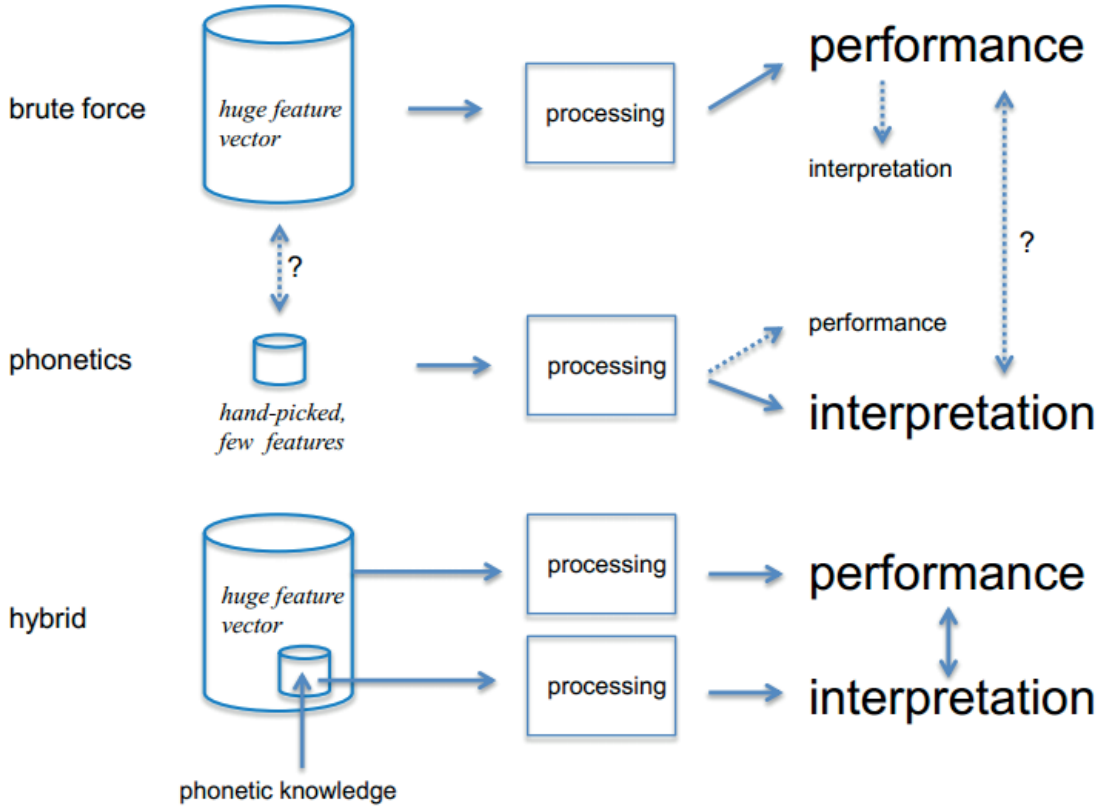

Abb. 5: Darstellung der im Projektverlauf getesteten hybriden Forschungsstrategie, die Performanz-Orientierung des brute-force Ansatzes mit der Theorie- und Interpretations-Orientierung des wissensbasierten phonetischen Ansatzes kombiniert (vgl. Schuller & Batliner, 2013).

Diskussion und Ausblick

Die wesentliche forschungsstrategische Wertschöpfung der Untersuchung lag in der Integration von theoriegeleiteten und interpretations-orientierten phonetischen Ansätzen mit datengeleiteten und Performanz-orientierten brute-force Ansätzen. Im Rahmen des Projekts wurde eine Lösung zur Harmonisierung von Performanz- und Interpretationsorientierten Forschungsstrategien entwickelt (vgl. Abbildung 5). Die ingenieurwissenschaftliche datengeleitete und Performanzorientierte („Getting Better") brute-force-Perspektive extrahiert große, hoch redundante Merkmalsvektoren, um sie in nachfolgenden datengeleiteten Merkmalsselektions- und Maschine-Learning-Verfahren weiterzuverarbeiten. Die Erkennungsleistung dieses Ansatzes ist optimiert, die Interpretation der relevanten phonetischen Prozesse bleibt jedoch größtenteils im Dunkeln. Diesem Ansatz stand bisher unvereinbar der Interpretationsorientierte phonetische Ansatz gegenüber, der mit kleinen, theoriebasierten, manuell selektierten Merkmalsvektoren und der Zielrichtung des Verstehens und Interpretierens der zugrundeliegenden Prozesse arbeitet („Getting Wiser"). In der hier umgesetzten integrativen Forschungsperspektive wurde ein hybrider Ansatz gewählt, der Performanz und auditiv-perzeptuelle Interpretierbarkeit kombiniert, indem die bisher wenig anschaulichen Merkmalsvektoren des brute-force Ansatzes um händisch selektierte Merkmale ergänzt und separat verarbeitet werden.

Einschränkung beziehen sich auf das experimentelle Setting in der die Sprechaufgaben stattfanden, hier wäre zugunsten der externen Validität die Wahl eines realitätsnäheren Settings mit unterschiedlichen Sprechkontexten eine Möglichkeit für zukünftige Forschungsbemühungen. Weitere Bemühungen beziehen sich auf den Aufbau einer multimodalen Sprach-, Video- und Biosignal-Datenbank zur Erfassung von Schläfrigkeit

Literatur

Burton, A. R., Rahman, K., Kadota, Y., Lloyd, A., & Vollmer-Conna, U. (2010). Reduced heart rate variability predicts poor sleep quality in a case–control study of chronic fatigue syndrome. Experimental Brain Research, 204(1), 71-78.

Honig, F., Batliner, A., Bocklet, T., Stemmer, G., Nöth, E., Schnieder, S., & Krajewski, J. (2014) Automatic analysis of sleepy speech. In ICASSP 2014, International Conference on Acoustics, Speech, and Signal Processing.

Kreiman, J., & Sidtis, D. (2011). Foundations of voice studies: An interdisciplinary approach to voice production and perception. John Wiley & Sons.

Schuller, B., & Batliner, A. (2014). Computational paralinguistics: emotion, affect and personality in speech and language processing. John Wiley & Sons.

Schuller, B., Batliner, A., Steidl, S., Schiel, F., & Krajewski, J. (2011). The interspeech 2011 speaker state challenge. In Proceedings INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, 3201-3204.

Schuller, B., Friedmann, F., & Eyben, F. (2014). The Munich Biovoice Corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production. In Language Resources and Evaluation Conference.

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. Language in society, 1(02), 179-195.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., & Pantic, M. (2014, November). Avec 2014: 3d dimensional affect and depression recognition challenge. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge,pp. 3-10. ACM.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., & Pantic, M. (2013). AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge,pp. 3-10. ACM.