

Akustische Stimmessung – Automatische Detektion von Befindlichkeitszuständen

Jarek KRAJEWSKI

Arbeits- und Organisationspsychologie, Universität Wuppertal
Gaußstr. 20, D-42097 Wuppertal

Kurzfassung: Potenzielle Anwendungsfelder der real-time Detektion von Befindlichkeitszuständen liegen in der Fahrerzustandserkennung, Unfallprävention, Mensch-Maschine-Interaktion, Psychotherapie und Eignungsdiagnostik. Die Analyse von Befindlichkeitszuständen erfordert in diesen Kontexten minimal-invasive, belästigungsfreie und den Tätigkeitsvollzug nicht beeinträchtigende Messzugänge. Ein vielversprechender neuer Messansatz für die real-time Detektion von internen Zuständen ist die akustische Stimmanalyse. Dieser Beitrag gibt einen Überblick zu Anwendungsfeldern, linguistisch-phonetischen Grundlagen, zentralen akustischen Beschreibungskategorien (Artikulation, Stimmqualität und Prosodie), Ablauf des Messprozederes und empirischen Validierungsbefunden. Die Erkennungsleistung liegt für Befindlichkeitszustände in 2-Klassenproblemen bei gegenwärtig 80% richtiger Zuordnungswahrscheinlichkeit. Zukünftige eignungsdiagnostisch-motivierte Forschungsbemühungen könnten sich über die Detektion von State Merkmalen hinaus, vor allem auf die stimmliche Erfassung von Traits (z.B. den Big Five) konzentrieren.

Schlüsselwörter: Stimme, automatische Zustandserkennung, ökologische Validität, Artikulation, Stimmqualität.

1 AUTOMATISCHE DETEKTION VON BEFINDLICHKEITZUSTÄNDEN

1.1 Anwendungsfelder

Die automatische real-time Erfassung von internen Zuständen aus sprachlichen Äußerungen besitzt in den Anwendungsfeldern (a) Fahrerzustandserkennung und Unfallprävention, (b) Human-Computer Interaktion, (c) marktpsychologische Produktbewertung, (d) Psychotherapie und (e) Eignungsdiagnostik vielversprechendes Einsatzpotenzial. Innerhalb von eignungsdiagnostischen Beratungs- oder Auswahl-situationen könnte eine aus State-Variablen abgeleitete, effiziente und vorurteilsfreie Trait-Messung von eignungsdiagnostisch relevanten Persönlichkeitsmerkmalen (z.B. Gewissenhaftigkeit, soziale Verträglichkeit, Integrität, Belastungssouveränität, Macht- und Leistungsmotiv) stattfinden. Neben Persönlichkeitsmerkmalen ist aber auch die Diagnostik von Kompetenzen und kognitiven Fähigkeiten - wie z.B. die Hochbegabendiagnostik mittels frühkindlicher Lautäußerungen - denkbar. Nicht zuletzt ergäben sich für psychologische Feldstudien die Möglichkeit in ökologisch validen Settings, belästigungsfrei und über einen längeren Untersuchungszeitraum Daten zu erheben (vgl. auch Fahrenberg, Myrtek, Pawlik & Perrez, 2007).

1.2 Messzugänge

Zur realtime Erfassung von Befindlichkeit stehen derzeit folgende Messansätze zur Verfügung: (a) Behaviorale Ansätze. Emotionale Informationen können aus dem Verfolgen von Gesichts-, Kopf-, Extremitäten- und Körperbewegung erschlossen werden. Typische Merkmale sind die räumliche Ausdehnung, Menge, Amplitude, Dauer, Flüssigkeit, Rhythmik und Dynamik der Bewegungen (Kollias et al., 2004). Weiteres Bewegungsverhalten wie die Mouse-, Keyboard- oder berührungssensitive Touchpad Interaktion (Schuller, 2006; Zimmermann et al., 2003) oder die Lenkrad- und Pedalbewegung bei der Fahrzeugbedienung könnten auch zur Detektion von Befindlichkeitszuständen herangezogen werden.

(b) Augenbewegungs-Ansätze. Zur Erfassung von mentaler Beanspruchung oder Schläfrigkeit Zuständen kommen gegenwärtig die Analyse von Augensakkadenbewegungen, Lidschlussverhalten - sowie Instabilitätsmaße der Pupillengröße zum Einsatz.

(c) Physiologische Ansätze. Das Messen physiologischer Daten wie EDA, EMG, EEG, EKG bietet vielfältige Möglichkeiten zur Bestimmung von Befindlichkeitszuständen.

1.3 Added-value des akustischen Messansatzes

Der Einsatz akustischer Informationskanäle zur Bestimmung des Befindlichkeitszustandes ist schon seit vielen Jahren anvisiert (Liebermann & Michaels, 1962). Aber erst die Fortschritte der Prozesstechnik und die Entwicklung leistungsfähiger Sprachanalyse-Software (z.B. Praat, Matlab) machten den breiten Einsatz von Sprachverarbeitung auf hohem Niveau möglich. Bausteine zur automatischen Bestimmung von internen Zuständen aus Stimmäußerungen (Speech Acoustics) liefern eine Reihe wissenschaftlicher Disziplinen: Künstliche Intelligenz Forschung, Computerlinguistik, Signal Processing, Audio Processing, Phonetik, Phonologie, Phoniatrie, Logopädie, Psychologie, Psycholinguistik und Linguistik.

Im Gegensatz zu physiologischen Ansätzen bietet der akustische Stimmmessungs-Ansatz, die Vorzüge eines berührungs- und sensorapplikationsfreien, belästigungs- und kalibrierungsarmen Messzugangs. Durch ein sprecherunspecifisches Vorhersagemodell, bzw. durch eine automatische Anpassung während der Bedienung (die erkannte Emotion dient als rückgekoppelte Überwachungsgröße) vermeidet der akustische Ansatz eine aus Komfortgründen unerwünschte Kalibrierungs- und Trainingsphase. Die Vorzüge des akustischen Ansatzes gegenüber behavioralen Ansätzen liegen darin, dass die Mikrophonaufnahme im Gegensatz zu der visuellen Informationsextraktion das Gefühl einer zu starken Beobachtung vermeidet. Auch ist die benötigte Standardhardware kostengünstig und in vielen neueren Fahrzeugsystemen bereits serienmäßig installiert. Für eine wachsende Verfügbarkeit verwertbarer Sprachäußerungen sorgen die in Zukunft vermehrt sprachgesteuerten Mensch-Maschine Schnittstellen, wie z.B. Spracherkennungstechnologie in Fahrerassistenzsystemen (z.B. Navigationssysteme, sprachgesteuerte Mobiltelefon Einwahl und Multimedia Anwendungen) und Computerarbeitsplätzen. In Folge dieser Entwicklung wächst der Bedarf an „emotional intelligenter“ Kommunikation, deren Effizienz und Akzeptanz von der realtime Detektion und adaptiven Reaktion auf momentane Anwender-Emotionen abhängt. Die Integration von akustischen Systemen in den Verbund einer übergeordneten multimodalen Systemarchitektur könnte wiederum den Nachteil temporär nicht verfügbarer Spracheingaben kompensieren und über eine wechselseitige Kreuzvalidierung der Informationskanäle die Akkuratheit und Robustheit erhöhen.

2 LINGUISTISCHE UND PHONETISCHE GRUNDLAGEN

2.1 Auditiv-perzeptive Beschreibungskategorien

Zur Beschreibung stimmlicher Veränderungen dienen die Klassen (a) Prosodie, (b) Artikulation und (c) Stimmqualität. Unter Prosodie fallen suprasegmentale (über mehrere Einzellaute verlaufende) Stimmphänomene wie Lautstärke, mittlere Sprechstimmlage, Sprechmelodie, Betonungswechsel, Sprechrhythmus, Sprechtempo und Sprechpausen. Artikulationsbezogene Beschreibungskategorien werden auf segmentaler Ebene betrachtet. Es sind Lippen- und Kieferbewegung (eingeschränkt vs. übermäßig), Zahnreihenabstand, Zungenbewegung (Vor- und Rückverlagerung der Zunge) und Lautbildung(sfehler) (verwaschen, unverständlich, Zusammenziehen von Worten). Die Stimmqualität schließlich wird über Merkmale des Stimmklangs (fest, klangvoll, wohlklingend, modulationsfähig, kräftig, dumpf), der Hyperfunktionalität (rau brüchig, knarrend, kratzend, gedrückt, gepresst, blechern, kloßig), der Hypofunktionalität (leise, matt, verhaucht, belegt, zittrig, klangarm) und der Stimm- und -absätze (weich, fest, hart, verhaucht) beschrieben. Eine zusätzliche Beobachtungskategorie bildet die Sprech- und Ruheatmung (angestrengt, schnappend, ziehend vs. mühelos, fließend, ausgeglichen). Gebräuchlich ist auch die gröbere Differenzierung in modale, falsetto, whispery (flüsternde), breathy

(behauchte), creaky (knarrende) und rough (raue) Stimmqualitäten. Darüber hinaus wird zur auditiv-perzeptuellen Beurteilung von Stimmqualität häufig das RBH System (Beurteilung nach Rauigkeit (R), Behauchtheit (B) und Heiserkeit (H) verwendet.

2.2 Akustische Stimmerkmalsklassen

Prosodiebezogene akustische Beschreibungsgrößen. Die oben genannten prosodischen Merkmale (Hauptgruppen: Intensität, Intonation und Dauer) sind die am häufigsten betrachteten akustischen Merkmale. Unter ihnen werden suprasegmentale, also über mehrere Einzellaute verlaufende, Eigenschaften der Melodie und des Rhythmus in der Sprechweise zusammengefasst.

- *Intensität.* Zur Bestimmung der Lautstärke wird die mittlere quadrierte Amplitude des Rohsignals innerhalb eines Zeitsegments berechnet. Wichtig ist bei der Messung von intensitätsbezogenen Lautstärke-Größen, dass Aufnahme- und Mikrofon-Distanz konstant gehalten werden (wie in Fahrzeugszenarien umgesetzt) oder nur Änderungen der Energie betrachtet werden.
- *Intonation.* Zur Bestimmung der Tonhöhe und Sprechmelodie wird der Verlauf der Fundamental-Frequenz (Vibrationsrate der Stimmlippen) berechnet.
- *Dauer.* Als Kenngrößen werden z.B. die Sprechrate (Anzahl der Sprachsegmente pro Zeit), Dauer von Pausen (innersilbisch, zwischen Phrasen), Anteil Pausen pro Sprechzeit sowie die Dauer von Vokalen bestimmt.

Artikulationsbezogene akustische Beschreibungsgrößen. Der artikulatorische Aufwand bei der Lautbildung von Konsonanten wird über die Berechnung des spektralen Schwerpunkts (Frequenzposition und spektraler Energiewert) genähert. Die aus dem Verlauf des spektralen Zentroids gebildeten Funktionale sind z.B. Langzeitmittelwert, Standardabweichung, relatives Maximum, mittlere betragsmäßige Steigung, Standardabweichung der betragsmäßigen Steigung und maximale betragsmäßige Steigung. Die Analyse der Vokalartikulation verwendet die Lage der ersten beiden Formanten, sowie ihren Bezug zu phonetischen Normwerten (Formantenpräzision) (Kienast & Sendlmeier, 2000).

Stimmqualitätsbezogene akustische Beschreibungsgrößen. Das Verhältnis ganzer spektraler Bänder, die Verhältnisse der Energien einzelner Harmonischer zur Gesamtenergie, die Regressionssteigung der Energieverteilung im Frequenzband über 1 kHz prägen u.a. den Stimmqualitätseindruck. Hinzu kommen kleine Schwankungen und Irregularitäten der Tonhöhe und Intensität (Jitter, Shimmer) sowie die Lage und Bandbreite von Formanten. So ist die Klangfarbe z.B. abhängig von der Teiltonstruktur sowie der Anzahl und der Stärke der im Klang enthaltenen Obertöne (eine große Anzahl an Obertönen indiziert tragfähige Stimmen). Energiekonzentrationen in hohen Frequenzen hingegen erzeugen einen hellen Stimmklang. Zu den wichtigsten Einzelkennzahlen zählen die Resonanzfrequenzen des Vokaltrakts (Maxima im Spektrum), die Formanten. Sie sind sensitiv für kleine Veränderungen der Vokaltraktform, wie sie durch ein Lächeln (Verkürzung des Vokaltrakts, Erhöhung der Formantenposition) oder eine gekrümmte Körperhaltung bewirkt werden. Weitere bedeutende Stimmqualitätsmaße sind Mel-Frequenz-Cepstral-Koeffizienten (MFCC), Roll-Off-Punkte, der spektrale Fluss, Harmonic-to-Noise Ratio und Hammarberg Indizes (vgl. Stevens & Hanson, 1994).

Zusammenfassend listet Tabelle 1 die gebräuchlichsten Kennzahlenfamilien auf und umreißt ihre Bedeutung. Die letztendlich berechneten Kennzahlen leiten sich aus den deskriptiven Kennwerten des jeweiligen zeitlichen Verlaufs der Beschreibungsgrößen ab: Minimum, Maximum, Median, Standardabweichung, Kurtosis, Schiefe, Steigung der Regressionsgerade, Regressionsfehler erster bis n-ter Ordnung und Mittelwert der ersten und zweiten Ableitung.

Tabelle 1: Akustische Beschreibungsgrößen aus den Stimmmerkmalsklassen Prosodie, Artikulation und Stimmqualität

Beschreibungsgröße	Erläuterung
<i>Prosodie</i> Intensität F0 F0 Contour Sprechrate	Mittlere quadrierte Amplitude innerhalb eines Zeitsegments („Lautstärke“) Fundamental Frequenz (Vibrationsrate der Stimmlippen) („Tonhöhe“) F0 Verlauf („Intonation“, „Sprechmelodie“) Anzahl Sprachsegmente pro Zeit
<i>Artikulation</i> Formant (F1-F2) Formant Präzision Spekt. Schwerpunkt	Resonanzfrequenzen des Vokaltrakts (Energie Konzentrationen im Frequenz Spektrum) Abweichung der Formantpositionen von phonologischen Normmaßen Zentroid der spektralen Energieverteilung (artikulatorischer Aufwand bei Konsonanten)
<i>Stimmqualität</i> Jitter/ Shimmer Spekt. Energie Distrib. Spekt. Fluss HNR Formant bandwidth MFCC	Kleine Schwankungen und Irregularitäten der Tonhöhe /Intensität Relative Stärke einzelner Frequenzbänder (z.B. HF 500) Veränderung der spekt. Energieverteilung pro Messfenster Harmonic-to-Noise Ratio (Anteil der periodischen Signalenergie) Breite des Resonanzpeaks (-3dB Schwelle) 12 Mel-Frequency Cepstrum Coefficients (sowie 12 Δ und 12 $\Delta\Delta$ MFCC)

3 OPERATIVER ABLAUF DES AKUSTISCHEN MESSPROZESSES

Der Messprozess untergliedert sich in 5 Verarbeitungsstufen: (a) *Sprachaufnahme*, (b) *Vorverarbeitung*, (c) *Merkmalsextraktion*, (d) *Merkmalsselektion* und (e) *Klassifikation*. Im Anschluss an die Sprachaufnahme folgt eine filternde und segmentierende Vorverarbeitung. Die anschließende Extraktion von prosodischer, stimmqualitativer und artikulatorischer Merkmalsverläufen wird von der Berechnung von statischen Beschreibungsgrößen der Merkmalsverläufe gefolgt. Die nachfolgende automatische Selektion und Reduktion von Merkmalen sind weitere wichtige Bestandteile zur Steigerung der Erkennensleistung. Der Einsatz von automatischen Klassifikatoren sowie ihre abschließende Kombination durch Metaklassifikatoren setzt den Schlusspunkt des Messprozederes.

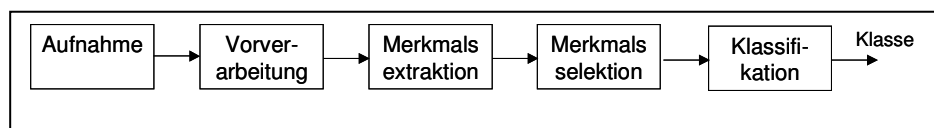


Abbildung 2: Darstellung des akustischen Messprozesses im Gesamtablauf

3.1 Sprachaufnahme

Das Sprachausgangsmaterial wird über eine Abtastfrequenz von 20 kHz mit 16 Bit Auflösung über einen Digitalrecorder aufgezeichnet. Weniger empfehlenswert, aber dennoch prinzipiell möglich, sind qualitativ schlechtere Aufnahmen wie sie beispielsweise aus Telephone oder Funkgesprächen resultieren. Kaum problematisch scheint hingegen die Verwendung von clip-on Mikrofonen oder Freisprecheinrichtungen. Bei der Auswahl des Sprachmaterial ist zwischen der Phonation einzelner Laute, dem Lesen von Silben-, Wort- und Satzlisten (z.B. Zungenbrecher oder phonetisch ausbalancierte Texte wie „Der Nordwind und die Sonne“), der Produktion von Wortreihen (z.B. Wochentage auflisten), freier (Bild-)Beschreibungen, monologischen/dialogischen Mensch-Maschine Eingaben und der monologischen/dialogischen Spontansprache zu unterscheiden. Dieses Sprachmaterial lässt sich wiederum mit diversen Sprechaufgaben (z.B. komfortabel, laut, euphorisch) kombinieren. Zu berücksich-

tigen ist bei der Auswahl des Formates jedoch, dass für ökologisch valide Studienergebnisse der Aufnahme und der spätere Anwendungskontext in Deckung gebracht werden sollten. Andererseits erleichtert eine Kontrolle über den gesprochenen Inhalt die Konstanzhaltung sprachinhaltsgebundener phonetischer Störgrößen.

Als Alternative zu der eigenständigen Aufnahme von Sprachmaterial kommt der Rückgriff auf Sprachdatenbanken in Frage. Sie beinhalten sowohl Sprachaufnahmen als auch die dazugehörigen validierten Emotionszuordnungen (Annotationen). Datenbanken mit emotionalen Sprachbeispielen sind bei Ververidis und Kotropoulos (2003) zu finden. Wichtige Datenbanken sind AEC, Sympafly, EMO-DB, IFA und DES. Güteigenschaften dieser Datenbanken bzw. des verwendeten Sprachmaterials sind: (a) Stichprobengröße (hohe Gesamtzahl an Lern- und Testbeispielen). In der Regel ist mit steigender Zahl an Lernbeispielen auch eine höhere Gesamtleistung eines Erkennungssystems auf Grund exakterer Modellierung der Problemstellung zu erwarten. Um die Bevorzugung einer überrepräsentierten Klasse zu vermeiden, sollten für alle Kategorien gleichmäßig verteilte Daten vorliegen; (b) ökologische Validität: Die Sprecheraufnahmen sollen der intendierten Zielpopulation und Anwendungssituation entsprechen. Eine angepasste Diversität bezüglich Altersklassen, Geschlecht, Bildungsgrad, kulturellen und sozialen Hintergründen vermeidet die Über- oder Unterschätzung der berechneten Zielbefindlichkeits-Detektionsraten; (c) Validität der Validierungsanker: Die zur Validierung häufig herangezogenen Perzeptionstests sollen über eine möglichst hohe Zahl von individuellen Annotatoren oder Experten abgesichert werden. Berücksichtigt werden sollte auch die Validitätsproblematik von geschau-spielten Emotionen.

3.2 Vorverarbeitung

Vor der Berechnung akustischer Kennzahlenverläufe steht in der Phase der Vorverarbeitung, die sich aus Zielsprecher Stimmaktivitätserkennung, Filterung und Segmentierung zusammensetzen kann.

- *Zielsprecher Stimmaktivitätserkennung.* Wenn spontane Sprachaufnahmen der Benutzerschnittstelle ohne manuelle Tasteninitiiierung im offenen Mikrofonbetrieb aufgezeichnet werden sollen, ist eine Stimmaktivitätserkennung erforderlich, um die Stimmaktivitätsphasen des Zielsprechers in dem kontinuierlichen Signalstrom von Nebengeräuschen und Fremdsprecheranteilen zu identifizieren. Die Eliminierung oder Verminderung von Störungen und Nebengeräuschen (z.B. Motor-, Wind und Fahrtgeräusche einer Autobahnfahrt) kann über die Optimierung der Signal-Noise-Ratios die Identifikationsleistung der Zielbefindlichkeit steigern.
- *Segmentierung.* Bevor die akustischen Kennzahlen bestimmt werden können, muss die Analysegranularität festgelegt werden. Welche sprachlichen Segmente wie Phoneme, Silben, Worte oder Intonationsphrasen (Äußerungen, Tunes) sind geeignet, um den Zielzustand bestmöglich zu erfassen. Das Selektieren der „units of interest“, der relevanten sprachlichen Abschnitte, kann manuell oder automatisch erfolgen. Bei der im Falle eines Echtzeitbetriebs erforderlichen automatischen Segmentierung wird die Untergliederung über die Detektion von Sprechpausen (Intonationsphrasen- Ebene) oder mit Hilfe von Automatic Speech Recognition Systemen (Phonem-, Silben-, Wort-Ebene) erzielt.

3.3 Merkmalsextraktion

Mittels Sprachanalysesoftware (z.B. Praat oder Matlab Toolbox) werden aus dem digitalisierten Rohsignal (Waveform) verschiedene Konturverläufe (siehe Kap. 2.2) aus den Bereichen Prosodie, Artikulation und Stimmqualität extrahiert. Typische prosodiebezogene Merkmalsverläufe sind: Intonation, Intensität und Dauer. Artikulationsbezogene Konturen beinhalten Formante und spektrale Zentroidverläufe. Stimmqualitätsbezogene Konturen sind z.B. Roll-off, Spektraler Fluss, MFCC, Jitter oder Shimmerverläufe. Nach dem ersten Schritt der Konturextraktion werden in einem Zweiten die eigentlichen Kennwerte gebildet. Hierzu werden Funktionalbildungen angewendet. Funktionaltypen sind lineare Momente (1.- 4. Ordnung, Mittelwert, Standardabweichung, Kurtosis, Schiefe) und weitere Beschreibungsgrößen wie 2.-4. Quartilposition, Maximum, Minimum, Range, Regressionsgradensteigerung, Regressionsfehler, Mittelwerte und Extrema der ersten und zweiten Ableitung. Neben der Be-

stimmung absoluter Extrema von Wertereihen bietet es sich unter Umständen an, diese zu relativieren. Dies geschieht mit Bezug auf den Mittelwert, und ist etwa bei der Erfassung des Intensitätsmaximums sinnvoll, um eine Normierung zu treffen. Darüber hinaus können auch die sprecherspezifisch z-normalisierte Kennwerte einer Verbesserung der Vorhersagekraft dienen.

3.4 Merkmalsselektion

Der Pool der potenziell relevanten Merkmale wird in einem nächsten Schritt auf diejenigen reduziert, die eine hohe Diskriminierungs-Leistung erzielen. Ziel ist es hierbei einerseits Rechenzeit bei der Merkmalsextraktion und Klassifikation zu reduzieren und andererseits die Gesamtzuordnungsleistung zu optimieren, indem der negative Einfluss irrelevanter Merkmale reduziert wird. So kann die zusätzliche Aufnahme nur einer Zufallsvariablen die Gesamtvorhersage-Leistung bereits um 10% verschlechtern. Werden hingegen zu viele Merkmale selektiert kommt es zu einem Informationsverlust und einer Verschlechterung der Klassifikationsleistung.

3.5 Klassifikationsverfahren

Das Ziel dieses wichtigen Phasenabschnitts ist es, die Vorhersageleistung einzelner Stimmindikatoren durch simultane Bündelung ihres Informationsgehaltes mittels Fusionierungsalgorithmen zu erhöhen. Die Verfahren der automatischen Mustererkennung ordnen nach einer Lernphase unbekannte Muster einer Klasse zu (z.B. über Lineare Diskriminanzanalysen). Die Muster setzen sich aus den selektierten Stimmmerkmalen zusammen, die zuzuordnenden Klassen entsprechen den vorherzusagenden Zielbefindlichkeiten. Für diese Aufgabe stehen eine Reihe von Klassifikatoren mit unterschiedlichen Stärken und Schwächen zur Auswahl. Als Anforderungen an ein Klassifikationsverfahren gelten: optimale Erkennungsleistung, Umgang mit fehlenden Werten, Trainingsstabilität, Toleranz gegenüber Dimensionserhöhung, Echtzeitfähigkeit in der Erkennung, kurze Trainingszeit, geringer Bedarf an Lernbeispielen, Rechenleistung und Speicher.

Einige der am häufigsten zur akustischen Emotionserkennung eingesetzten Klassifikationsverfahren sind: Lineare Diskriminanzanalyse, Naive-Bayes, Künstliche Neuronale Netze, Support-Vektor-Maschinen, K-Nearest Neighbour Verfahren, Bayessche Netze, Dynamic-Time-Warping (Schuller, 2002d), Entscheidungsbäume, Gaußsche-Mixtur-Modelle und Hidden-Markov-Modelle.

Verbessert werden kann die Klassifikationsleistung (ca. + 5% bis 10% Zuordnungsgenauigkeit) durch eine Kombination der Klassifikationsverfahren in multiplen Instanzen (Ensemble-Klassifikation). Die Klassifikationsergebnisse der Level-0-Klassifikatoren werden in einem nachfolgenden Schritt mittels einer finalen Instanz, dem übergeordneten Level-1-Klassifikator verarbeitet. Im einfachsten Fall der Kombination der Level-0 Ergebnisse, dem Majority Voting, werden die endgültigen Klassenzuordnungsentscheidungen über einen Mehrheitsentscheid gefällt (Voting). Darüber hinaus lassen sich die Ensemble Klassifikatoren z.B. mittels Stacking und Boosting kombinieren (Schuller, 2006).

3.6 Evaluierungsstrategien

Die in der psychologischen Forschung überwiegend angewandte Validierungsstrategie bestimmt an einem ungespliteten Datensatz sowohl die Modell-Parameter (z.B. b-Regressionsgewichte) als auch die Modellanpassungsgüte (z.B. R^2). Dieses Vorgehen führt (trotz Korrekturversuchen wie dem „korrigierten R^2 “) zu Überschätzungen der tatsächlichen Vorhersagekraft der Modelle. Um eine realistische Bestimmung der Erkennungsleistung abzuleiten, muss daher in einem Trainingsdatensatz das Modell trainiert werden und in einem davon disjunkten Datensatz die Anpassungsgüte getestet werden.

Da dieses Prozedere die zum Training des Modells zur Verfügung stehende Datenmenge reduziert, kommen stichprobeneffiziente Vorgehen wie die Leave-One-Speaker-Out (LOSO) Strategie oder die j-fach stratifizierte Kreuzvalidierung (SCV) zum Einsatz (Witten & Frank, 2000). Dabei wird die gesamte Datenmenge zuerst in j getrennte Teilmengen partitioniert. Anschließend werden in j Testdurch-

läufen jeweils die j -te Teilmenge als Testmenge und die $j-1$ verbleibenden Teilmengen als Trainingsmengen verwendet. Als Gesamterkennungsleistung wird der Mittelwert aus den Einzelerkennungsleistungen der j Durchläufe berechnet. Im einfachsten Fall wird der Datensatz in zwei 50% umfassende Datenpartitionen gesplittet und die Teilmenge 1 einmal als Trainingsmenge und einmal als Testmenge verwendet. Andere gebräuchliche Partitionierungen nutzen Trainingsdatenanteile von 66,7%, 80%, 90% oder 95%. Für die Bewertung der erzielten Erkennungsleistungen stehen Kenngrößen wie Detektionsrate (Anteil der z.B. als müde klassifizierten, müden Sprecher), False Alarm Rate (Anteil der als müde klassifizierten, nicht-müden Sprecher), overall Recognition Rate (RR; richtig zugeordneten Sprecher durch Gesamtanzahl der Zuordnungen), klassenweise gemittelte Recognition Rate (CL; klassenweise, unabhängig von der Klassengröße, gemittelte Detektionsraten). Bei der Bewertung der erzielten Klassifikationsleistungen sollten interne und externe Validitätsüberlegungen eine Rolle spielen. Für die Einschätzung der erzielten Detektionsraten sollten insbesondere folgende Fragen beantwortet werden:

- Gibt es eine an die anvisierte Anwendungssituation angepasste Diversivität hinsichtlich der Zusammensetzung der Zielpopulation? (allgemeine Sprechermerkmale wie z.B. Alter, Geschlecht und Dialekt)
- Gibt es eine an die anvisierte Anwendungssituation angepasste Diversivität hinsichtlich des Zustandsspektrums der Zielpopulation? (sind alle Zustände -und nur die- repräsentativ erfasst, zwischen denen das System differenzieren können soll?)
- Sind zu erwartende Zustands-Kombinationen erfasst, die die Detektion des einzelnen Ziel-Zustands maskieren könnten? (z.B. müde und ängstlich, müde und gereizt)
- Besteht Übereinstimmung in den gemachten Sprechersituationsannahmen (sprecherbezogene Trainingsmöglichkeiten, keine Sprecherüberlappung) und der realen Anwendungssituation?

4 EMPIRISCHE VALIDIERUNGSBEFUNDE

Zahlreiche Studien dokumentieren die Validität der akustischen Stimmanalyse in den Bereichen: Detektion von Basisemotionen, User States (z.B. Frustration, Belästigung), Stress, Alkoholintoxikation und Schläfrigkeit (z.B. Batliner et al., 2005; Cowie et al., 2001; Juslin & Laukka, 2003; Paeschke, 2003; Picard, 1997; Kienast & Sendlmeier, 2000; Klasmeyer & Sendlmeier, 2000; Krajewski et al., 2007; Scherer, 2001; Schuller, 2006; Zhou, Hansen & Kaiser, 1999). Die durchschnittliche Zuordnungswahrscheinlichkeit der prognostizierten Phänomene liegt für 2-Klassenprobleme jeweils zwischen 60-80% (entspricht einer Validitätskorrelation von ca. .5-.6) und einer Zuordnungswahrscheinlichkeit für 4-Klassenprobleme von 50-60%.

Basisemotion. Die überwiegende Anzahl an Arbeiten im Bereich der akustischen Stimmanalyse beschäftigt sich mit der Detektion von Basisemotionen (z.B. Batliner et al., 2005; Cowie et al., 2001; Paeschke, 2003; Kienast & Sendlmeier, 2000; Scherer, 2001). Exemplarische Erkennungsraten sind für fünf Emotionen 76,2% (Schuller, 2006) oder 77,4% (Vogt & Andre, 2005) und 88,8% (Schuller, 2006; vgl. auch Tabelle 2) bei sieben Emotionen. Im Gegensatz zu den hier verwendeten gespielten und unter Laborbedingungen aufgezeichneten Sprachdaten liegen die Ergebnisse unter realistischen Bedingungen von spontanen, im Feld aufgezeichneten Emotionen der Datenbank AEC bei 79,1%.

Tabelle 2: Vergleich Erkennungsleistung Mensch-Maschine nach Emotionen, Datenbank EMO-DB, Klassifikation mit Support Vector Maschine, 10-fach stratifizierte Kreuzvalidierung (Schuller, 2006)

Recognition Rate [%]	Ärger	Ekel	Furcht	Freude	Neutral	Langeweile	Trauer
Mensch	73,3	63,3	96,7	93,3	99,9	90,0	73,3
SVM-Klassifikation	92,9	92,1	89,1	69,0	84,6	89,9	92,5

Aktivierungsdimension der Emotion. Folgende stimmlichen Veränderungen ergeben sich auf Einzelkennzahlenebene für die emotionalen Aktivierungsdimension: niedrige Aktivierungszustände produzieren tiefe F0 Werte (Tonhöhe), flache Intonationsverläufe, niedrige Intensitäten, geringe Intensitätsschwankungen, niedrige erste Formant (F1) Positionen, enge F1 Bandwidth, niedrige F1 Präzisionen (verschliffene Artikulation) sowie geringe spektrale Energie der Frequenzen über 500Hz (Juslin & Laukka 2001; Scherer, 1986).

Schläfrigkeit. Schlafdeprivationsstudien geben Hinweise auf schläfrigkeitsinduzierte sprachliche Veränderungen wie eine verwaschene Aussprache, verlängerte Wortdauer, flache Intonation, tiefe F0 sowie eine geringe spektrale Hochfrequenzenergie (vgl. auch Harrison & Horne 1997; Whitmore & Fisher 1996). Die Korrelationen der wichtigsten Einzelstimmerkmale mit subjektiven Schläfrigkeitsscores bewegen sich in einem korrelativen Bereich von .2 bis .3 (Krajewski et al., 2007). Die mittels Linearer Diskriminanzanalyse bestimmte Vorhersageakkuratheit von Schläfrigkeit aus einer simulierten Fahrerassistenzsystem Eingabe („Ich suche die Friesenstraße“) beträgt 88.2% (Krajewski & Kröger, 2007).

5 RESÜMIERENDE SCHLUSSBETRACHTUNG

Eignungsdiagnostische Relevanz. Die real-time Erfassung von Befindlichkeitszuständen eröffnet eine Reihe von vielversprechenden Anwendungsfeldern. Die Anwendungskontexte erfordern jedoch einen den Tätigkeitsvollzug nicht beeinträchtigenden, belästigungsarmen Messzugang, wie den der akustischen Stimmessung. Durch die Berechnung von prosodischen, artikulatorischen und stimmqualitätsbezogenen Kennzahlen werden für eine Reihe von State-Merkmalen (z.B. Basisemotionen, Stress, Alkoholisiertheit und Müdigkeit) bereits heute unter realistischen Aufnahmesituationen gute bis sehr gute Klassifikationsgüten (ca. 80% richtige Zuordnungen) erzielt. Die ökologische Validität und Implementierbarkeit der Stimmessung in Alltagskontexte (vgl. Fahrenberg, Myrtek, Pawlik & Perrez, 2007) unterstreicht daher zusätzlich ihre potenzielle eignungsdiagnostische Bedeutung. Um dieses Potenzial zu entfalten und prognostisch relevante Auswahlinstrumente zu entwickeln, sollte in Zukunft daher der Fokus von State- zu Trait-Merkmalen wechseln. Vielversprechend wäre z.B. die (unbemerkt) stimmliche Messung von Big Five Variablen wie der Extraversion oder der Verträglichkeit.

Forschungsdessiderate. Zukünftige Forschungsbemühungen sollten sich entlang des 5-stufigen Messprozederes mit einer leistungsfähigeren Vorverarbeitung, Merkmalsextraktion, Merkmalsselektion, Klassifikation und Metaklassifikation beschäftigen. Darüber hinaus ist der Aufbau einer forschungsförderlichen Infrastruktur von großer Bedeutung. Dazu zählen frei zugängliche Programm-Skripte für die Bestimmung von Sprachkennzahlen und frei zugängliche Sprachdatenbanken für Befindlichkeitszustände. Diese Sprachkorpora sollten über die gewöhnlichen Standards hinaus auch Zustands-Kombinationen beinhalten (wie z.B. ein müder Sprecher der zugleich ängstlich, erkältet, gestresst ist, Alkohol/ Milch getrunken hat oder unter Schmerzeinfluss steht). Einen zusätzlichen Robustheits- und Leistungsgewinn verspricht die *multimodale Integration* von akustischen, visuellen, physiologischen und behavioralen Maßen. Der sich daraus ableitende Bedeutungszuwachs interdisziplinärer Forschungsk Kooperationen eröffnet der psychologischen Forschung wichtige zukunfts-fähige Beschäftigungsfelder.

LITERATUR

- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. & Taylor, J.G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18, 32-80.
- Fahrenberg, J., Myrtek, M., Pawlik, K. & Perrez, M. (2007). Ambulantes Assessment - Verhalten im Alltagskontext erfassen: Eine verhaltenswissenschaftliche Herausforderung an die Psychologie. *Psychologische Rundschau*, 58, 12-24.
- Fernandez, R. & Picard, R. (2002). Modeling drivers' speech under stress. *Speech Communication*, 40, 145-159.
- Harrison, Y. & Horne, J.A. (1997). Sleep deprivation affects speech. *Sleep*, 20, 871-877.
- Juslin, P.N. & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129, 770-814.
- Kienast, M. & Sendlmeier, W.F. (2000). Acoustical analysis of spectral and temporal changes in emotional speech. In *SpeechEmotion-2000*, 92-97.
- Klasmeyer, G. & Sendlmeier, W. F. (2000). Voice and emotional states. In R.D. Kent, & M.J. Ball (Eds.), *Voice Quality Measurement* (pp. 339-357). San Diego: Singular.
- Kollias, S., Amir, N., Kim, J. & Grandjean, D. (2004). *Description of potential exemplars: Signals and signs of emotion*. Report HUMAINE Human-Machine Interaction Network on Emotions. Athens: Univ. Press.
- Krajewski, J. & Kröger, B. (2007). Using Prosodic and Spectral Characteristics for Sleepiness Detection. *Proceedings Interspeech*, 94-98.
- Krajewski, J., Gundel, A., Wilhelm, B. & Kröger, B. (2007). Akustische Schläfrigkeitsmessung. In K. Jenewein (Hrsg.), *Kompetenzentwicklung in realen und virtuellen Arbeitssystemen*. Dortmund: GfA-Press.
- Lieberman, P. & Michaels, S. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *J. Acoust. SOC. Am.*, 34, 922-927.
- Paeschke, A. (2003). *Prosodische Analyse emotionaler Sprechweise*. Berlin: Logos.
- Picard, R.W. (2000). Towards computers that recognize and respond to user emotion. *IBM Systems Journal*, 39, 705-719.
- Scherer, K.R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99, 143-165.
- Scherer, K.R., Johnston, T. & Klasmeyer, G. (2003). Vocal expression of emotion. In R.J. Davidson, K.R. Scherer, H.H. Goldsmith (Eds). *Handbook of affective sciences*, 433-456.
- Schuller, B. (2006). *Automatische Emotionserkennung aus sprachlicher und manueller Interaktion*. Technische Universität München.
- Stevens, K. & Hanson, H. (1994). Classification of glottal vibration from acoustic measurements. *Vocal Fold Physiology*, 147-170.
- Ververidis, D. & Kotropoulos, C. (2003). A state of the art review on emotional speech databases. *Tagungsband Ist Richmedia Conference*, 109-119.
- Vogt, T. & Andre, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. *Tagungsband ICME 2005, 6th International Conference on Multimedia and Expo, IEEE*, 474- 477.
- Whitmore, J. & Fisher, S. (1996). Speech during sustained operations. *Speech Communication*, 20, 55-70.
- Witten, I.H. & Frank, E. (2000). *Data Mining: Practical machine learning tools with Java implementations*. San Francisco: Morgan Kaufmann.
- Zhou, G., Hansen J.H.L. & Kaiser J.F. (1999). *Methods for stress classification: Nonlinear teo and linear speech based features*. ICASSP '99, Phoenix (Arizona), 2087-2090.
- Zimmermann, P., Guttormsen, S., Danuser, B. & Gomez, P. (2003). Affective computing - A rationale for measuring mood with mouse and keyboard. *International Journal of Occupational Safety and Ergonomics*, 9, 539-551.