Sprachliche Emotionserkennung im Fahrzeug

Dipl.-Ing. Univ. Björn Schuller
Univ.-Prof. Dr.-Ing. habil. Gerhard Rigoll
Univ.-Prof. Dr. rer. nat. Manfred Lang
Lehrstuhl für Mensch-Maschine-Kommunikation
Technische Universität München
(schuller | rigoll | lang)@ei.tum.de

Zusammenfassung

In diesem Beitrag stellen wir sprachliche Emotionserkennung im Fahrzeug als Grundlage für eine emotionale Mensch-Maschine-Interaktion vor. Dem Boardcomputer soll mit dem Wissen über die aktuelle Fahreremotion die Möglichkeit verliehen werden sozial kompetent entscheiden zu können. Die Erkennung von Gereiztheit oder Irritation des Fahrers und Zuständen wie Trunkenheit oder Müdigkeit sowie weiteren mentalen Zuständen soll zur Einleitung von Sicherheits- und Fehlerauflösungsstrategien beitragen. Dabei wird sowohl das akustische Signal selbst sowie der gesprochene Inhalt einer natürlichsprachlichen Fahreräußerung betrachtet. In der prosodischen Analyse werden eine Reihe unterschiedlicher Merkmale und Klassifikationsverfahren vorgestellt und verglichen. Darüber hinaus wird der gesprochene Inhalt einer Fahreräußerung betrachtet und in das Ergebnis integriert. Eine für Training und Evaluation der stochastischen Modelle verwendete Sprachdatenbank und ihre Akquisition wird im Detail vorgestellt. Die Problematik der robusten Erkennung in der Kommunikation wird vor dem Hintergrund der Fahrzeugumgebung diskutiert. Schließlich werden Ergebnisse, die mit den besprochenen Methoden erzielt werden konnten, präsentiert und diskutiert.

1. Einleitung

Im Folgenden soll der Einsatz von Emotionserkennung als Stütze für die Entscheidung im Fahrzeug motiviert werden. Die Einteilung von Emotionen in dieser Domäne wird besprochen, und Sprache als Modalität für eine robuste Erkennung vorgestellt.

1.1 Anwendung im Fahrzeug

Folgende Anwendungsszenarien stehen im Vordergrund: Freude als Bezugsgröße für ein unüberwachtes Nachtrainieren hinsichtlich einer Adaptivität stochastischer Modelle im Fahrzeug. Dabei wartet das System ähnlich einem Schüler oder Diener die emotionale Reaktion eines Anwenders nach seiner Aktion ab. Fällt diese positiv aus, kann das

Quelle: DGLR-Bericht 2003-04, Universität der Bundeswehr, München

System sein Verhalten als richtig lernen. Im Gegenzug hierzu kann sich das System bei einem verärgerten Benutzer einerseits entschuldigen, oder gezielte Rückfragedialoge zur Aufklärung eines mutmaßlichen Fehlers aktiv einleiten. Ein irritierter Benutzer kann Anlass zur aktiven Hilfestellung geben. Auf der anderen Seite wird diese Hilfe so nur gezielt dargeboten. Als weitere Benutzerzustände werden die Müdigkeit oder Trunkenheit des Fahrers zur Einleitung von Sicherheitsstrategien genutzt.

1.2 Einteilung der Emotionen

In der automatischen Klassifikation von Emotionen haben sich vor allem mehrdimensionale Emotionsräume und der Einsatz von diskreten emotionalen Zuständen als Klassifikationsschemata durchgesetzt. In der vorgestellten Arbeit werden diskrete Emotionen betrachtet. Ursprünglich wurde ausgehend von einem eindimensionalen Emotionsraum mit der Achse Positivität auf einen zweidimensionalen Raum mit der weiteren Achse Aktivität übergegangen. Diese Darstellung erwies sich jedoch als zu komplex hinsichtlich der Erkennungsleistung in Bezug auf die Anwendbarkeit der erhaltenen Information. Zur besseren Vergleichbarkeit werden die Ergebnisse dieser Arbeit mit den Emotionen nach dem MPEG-Standard plus einem neutralem Zustand zur Abgrenzung präsentiert. Die in unseren Fahrzeug-Setups eingesetzten Emotionen bewegen sich in der Erkennungsleistung auf gleichem Niveau. Dabei wird die Zahl erkannter Emotionen jeweils bewusst auf die in der Applikation umgesetzten reduziert um eine sicherere Erkennung zu begünstigen.

1.3 Sprache als Modalität

Eine Reihe unterschiedlicher Modalitäten wird für die Erkennung von Benutzeremotionen eingesetzt. Für den Anwender weniger angenehm, jedoch relativ akkurat bieten sich traditionell invasive Verfahren wie die Messung von Herzfrequenz, Temperatur, Hautwiderstand, Feuchtigkeit, etc. an. Daneben existiert eine Reihe neuerer Ansätze wie bild- oder sprachbasierter Verfahren [02sch1] zur Erkennung aus der Mimik oder sprachlichen Äußerungen. Als berührungssensitives Verfahren konnte darüber hinaus auch die Erfassung über die Interaktion auf einem Touchscreen im Fahrzeug erfolgreich vorgestellt werden [02sch2]. Als besonders viel versprechend zeichnet sich der Einsatz kombinierter multimodaler Ansätze unter Einbeziehung von Kontextwissen ab. Im Fahrzeug kann hier etwa das aktuelle Fahrverhalten und die Interaktion mit dem MMI mit betrachtet werden. In diesem Beitrag konzentrieren wir uns jedoch auf die Erkennung aus dem Sprachsignal.

Der Einsatz von sprachlicher Emotionserkennung wird speziell im Fahrzeug durch den nicht-invasiven Charakter und die Gebräuchlichkeit

sprachlicher Äußerungen zur Emotionstransmission motiviert. Sinnvoll erscheint dies vor allem vor dem Hintergrund eines MMIs welches mit natürlicher Sprache vom Fahrzeugführer bedient wird. Der Benutzer hat als besonderen Vorteil über den Sprachkanal die Kontrolle, wie viel Emotion er zeigen möchte. Dies kann besonders bei invasiven Verfahren nicht gewährleistet werden. Als Aufnehmer kann für die automatische Erkennung ein standardmäßig vorhandenes Mikrofon verwendet werden. Die Emotion soll dabei aus dem semantischen Inhalt (vgl. [03dev1],[01lee1]) von Benutzeräußerungen sowie aus der prosodischen Charakteristik des Sprachsignals bestimmt werden. Dies erlaubt eine robustere Schätzung, auch vor dem Hintergrund ironischer Äußerungen. Darüber hinaus ist eine emotionale Interpretation einer insgesamt höheren Zahl an Benutzeräußerungen möglich, da nicht alle lautsprachlichen Äußerungen semantisch verwertbaren Inhalt hinsichtlich der Emotion besitzen. Als Parameter für die Optimierung einer Erkennung sind dabei im Wesentlichen die Aufnahmetechnik, die Vorverarbeitung, die verwendeten Merkmale, das gewählte Klassifikationsverfahren und eine sinnvolle Integration unter Einbezug von Konfidenzen zu nennen. In diesem Beitrag beschränken wir uns bei der Optimierung auf die Wahl der Merkmale und das Klassifikationsverfahren. Ziel einer optimierten Erkennung wird dabei die Unabhängigkeit vom Sprecher, der Sprache, speziell bei der akustischen Erkennung vom gesprochenen Inhalt und schließlich eine kontinuierliche Erkennung auch im Störsignal sein.

2. Sprachkorpora

Um stochastische Erkenner für die automatische Erkennung von Emotionen trainieren und evaluieren zu können, müssen Beispiele sprachlicher Emotionsäußerungen zur Verfügung gestellt werden können.

2.1 Anforderungen an den emotionalen Sprach-Korpus

Zunächst sollte es sich um möglichst spontane und realistische Emotionsäußerungen handeln. Zur Erzielung hoher Erkennungsraten und größtmöglicher Unabhängigkeit vom späteren Anwender ist darüber hinaus eine hohe Gesamtzahl von Trainingsbeispielen möglichst vieler unterschiedlicher Sprecher mit gleicher Verteilung unter den Emotionen anzustreben. Die Sprecher sollten beiden Geschlechtern, verschiedenen Altersklassen, Bildungsgraden und Kulturen angehören. Um weiter auch von der Sprache, dem gesprochenen Inhalt und der Länge der Äußerung unabhängig zu werden, ist auch hier Diversität zu gewährleisten. Die Aufzeichnungen sollten des Weiteren von Störgeräuschen frei sein, wobei unter Umständen auch eine gezielte Aufzeichnung im späteren Störumfeld sinnvoll sein kann. Ein Problem bei der Erfassung von Trainingssamples ist die Zuordnung zu Emotionen, die möglichst der

tatsächlich empfundenen Emotion des Probanden entsprechen sollte. Generell lässt sich sagen, dass Beispiele des expliziten Benutzers aus dem potentiellen Umfeld und Verlauf ideal erscheinen. Ein standardisierter Korpus hingegen bietet die Möglichkeit internationaler Vergleichbarkeit der angewandten Verfahren, und ist besonders in dieser Domäne wünschenswert.

2.2 Akquisition emotionaler Sprachsamples

Prinzipiell bietet sich die Möglichkeit emotionale Sprachsamples aus gezielten Laborversuchen, Langzeitbeobachtungen, oder verdeckten Beobachtungen zu sammeln. Als weitere Möglichkeit werden gerne Sprach-Beispiele aus den Medien verwendet. Dabei lässt sich generell eine Einteilung zwischen spontanen Emotionen und gespielten Emotionen vollziehen.

Um eine Sprachdatenbank nach den genannten Kriterien anzulegen wurden zunächst eine Reihe von Laborversuchen zur Akquisition spontaner Emotionen vollzogen. Hierfür wurden Versuche mit gezielten Provokationen wie Beurteilungen durch den Versuchsleiter, Reizung unter Umständen auch weiterer Wahrnehmungskanäle, gestaffelten Gratifikationen, oder ähnlichem in typischen Fahrzeug-Bediensituationen vollzogen. Hierbei kam in einem Wizard-of-Oz orientierten Test-Setup eine semiautomatische Ablaufsteuerung [02nie1],[02sch4] und eine Fahrsimulation [02mcg1] für die Tests zum Einsatz, um einen effizienten Ablauf und konstante Bedingungen zu gewährleisten. Weiterhin wurde Spiele zwischen Probanden vollzogen, um möglichst spontane Emotionen zu erzielen. Als Vorteil bietet sich die Möglichkeit Probanden nach ihrem subjektivem Empfinden zu befragen zu sehen. Jedoch ist die Versuchsatmosphäre in der Regel präsent. Als Grenzfall spontaner Emotionen ist das Vorlesen emotionaler Texte durch den Probanden zu sehen. Als weitere Möglichkeit wurden Langzeitbeobachtungen vollzogen, in denen der Proband kontinuierlich im Alltag aufgezeichnet wurde. Ziel ist es hier, den Beobachtungseffekt zu reduzieren. Jedoch entsteht hierbei ein besonders hoher Überschuss an Daten, wobei mit Hilfe einer automatischen Segmentierung zumindest eine Einschränkung auf tatsächliche Äußerungen möglich ist. Die Zuordenbarkeit ist hierbei erschwert, da der Proband erst nach einer längeren Zeit zu seiner Emotion befragt werden kann. Zusätzlich ergibt sich zunächst eine verzerrte Verteilung unter den Emotionen, was über einen längeren Zeitraum gesehen sich jedoch der tatsächlichen Auftrittswahrscheinlichkeit annähert. Eine optimale Vermeidung von Beobachtungseinflüssen lässt sich prinzipiell ausschließlich durch eine verdeckte Beobachtung erreichen. Die hiermit größte Spontaneität wird jedoch nur auf Kosten geringerer Zuordenbarkeit, einen ebenfalls hohen

Überschuss an unbrauchbaren Daten, und in der Regel mit hohem additiven Störgeräusch erzielt. Nicht zuletzt auf Grund der rechtlichen Situation wurde auf diese Methode bewusst verzichtet. Ähnlich verhält sich dies bei Ausschnitten aus Medienbeiträgen. Auch hier ist ein hoher Aufwand bei der Segmentierung zu betreiben und in der Regel Störgeräusch präsent. Die Emotionen sind je nach Beitrag spontan oder gespielt, können aber nur aus der Situation von Dritten eingeteilt werden. Bei den gespielten Emotionen ist hier zumindest von höherer Professionalität auszugehen.

In einer zweiten Phase wurden ausschließlich gespielte Emotionen gesammelt. Diese bieten vor allem die Vorteile eindeutiger Zuordenbarkeit, Aufnahmen in Studioqualität, eine hohe Zahl erzielbarer Samples, und eine nach Wunsch gleiche Verteilung unter den Emotionen. Dabei können verschiedene Sprechertypen unter all den genannten Vorteilen erfasst werden. Somit bietet diese Möglichkeit eine sinnvolle und allgemein gerne verwendete Grundlage um zu einem großen Schatz an Beispielen zu gelangen, die in jedem Fall einen Test auf grundsätzliche Eignung gewählter Verfahren erlauben. Als Nachteile sind die Gefahr der übertriebenen oder verzerrten Darstellung der Äußerungen, sowie in Einzelfällen auch die mangelnde Ausprägung je nach schauspielerischer Begabung der Akteure zu nennen. Es sind Aufzeichnungen sowohl von Laien als auch von erfahrenen Darstellern im gesammelten Korpus enthalten. Die Emotionen wurden direkt und über einen größeren Zeitraum zur Vermeidung eines Lerneffekts jeweils nach einer Einübungsphase gespielt. Die Aufzeichnungen erfolgten in einem reflexionsarmen Raum unter Verwendung eines aktiven Kondensatormikrophons. Insgesamt wurden Daten von 38 Sprechern in den Sprachen Deutsch, Englisch und Mandarin Chinesisch erfasst.

3. Semantische Emotionserkennung

Im Folgenden soll die Erkennung aus dem gesprochenen Inhalt auf Basis sogenannter emotionaler Marker beschrieben werden. Oft ist der emotionale Zustand direkt aus dem gesprochenen Inhalt ableitbar. Im Fall von Ironie kann dies zu Missinterpretationen führen. Dies wird jedoch in der später diskutierten Fusion mit dem prosodischen Modell berücksichtigt. In Versuchen in der Fahrzeugdomäne lag der beobachtete Anteil emotionaler Phrasen der Probanden bei 5%. Hieraus resultiert ein hoher Überschuss aus nicht emotional gefärbten Phrasen, was einen spotting-basierten Ansatz motiviert. Die Basis der Interpretation bilden die n Hypothesen mit Einzelwortkonfidenz eines Spracherkenners auf HMM-Basis mit Zero-Grammen als Sprachmodell. Dabei haben sich in unseren Versuchen zehn Hypothesen als Optimum heraus kristallisiert. Die Emotionen werden hier als Benutzerintentionen

im Sinne einer intendierten Emotionsäußerung verstanden. Reine Wortmodelle auf Basis des Spracherkenners reichen nach unserer Meinung nicht für die Schätzung der Emotion aus, da auch Negationen eines Ausdrucks oder Angaben des Benutzers über die Ausprägung der Emotion verstanden werden sollen. Parallel zum Sprachverstehen in der Interaktion mit dem Board-MMI wird so ein emotionales Phrasenmodell evaluiert. Für das Spotting werden Bayes'sche Netze eingesetzt. Sie sind in der Lage unvollständige und unsichere Information zu behandeln und erlauben eine a-priori Gewichtung von Emotionen. Darüber hinaus bietet sich durch Erweiterung der klassischen harten Evidenzen hin zu weichen Evidenzen die Möglichkeit Einzelwort-Wahrscheinlichkeiten des Spracherkenners zu integrieren. In vier Ebenen erfolgt eine Klusterung von Beobachtungen über semantische Untereinheiten über Teil-Phrasen hin zu Intentionen. Parameter wie die Ausprägung der Emotion werden mit Konfidenz ausgegeben. Die bedingten Wahrscheinlichkeiten des Netzes werden aus Verschriftungen der Benutzerphrasen aus der Sprachdatenbank trainiert.

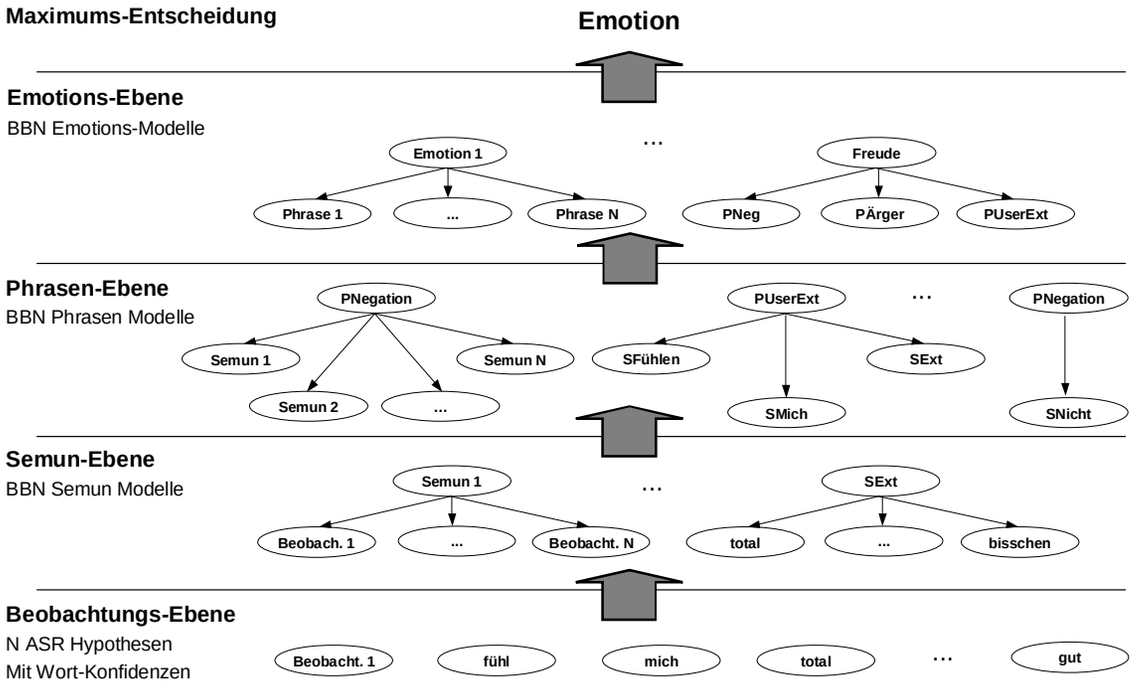

Abbildung 1: Bayes'sche Netze zum Spotting emotionaler Phrasen

Jede Emotion ist in einem eigenen Netz modelliert, wobei für das Netz mit der höchsten Wurzelwahrscheinlichkeit entschieden wird.

4. Prosodische Emotionserkennung

„Das Verständliche an der Sprache ist nicht das Wort selber, sondern Ton, Stärke, Modulation, Tempo, mit denen eine Reihe von Worten gesprochen wird - kurz die Musik hinter den Worten, die Leidenschaft hinter dieser Musik, die Person hinter dieser Leidenschaft: alles das also, was nicht geschrieben werden kann“ [Friedrich Nietzsche].

Im Folgenden soll die Analyse des Sprachsignals auf akustischer Ebene zur Erfassung der Fahreremotion im Detail besprochen werden. Dabei wird vor allem auf die Wahl geeigneter Merkmale und Klassifikatoren näher eingegangen.

4.1 Statische Merkmale auf akustischer Ebene

In der vorgestellten Arbeit kamen ausschließlich statische Merkmale zum Einsatz. Hierbei handelt es sich im Wesentlichen um abgeleitete Statistiken aus den Verläufen von low-level Features. In vorhergehenden Arbeiten [02sch4] wurde der Einsatz dynamischer Verfahren wie Hidden-Markov-Modelle zur Analyse der vollständigen Konturen evaluiert. In unseren Arbeiten [03sch1] erwiesen sich jedoch die abgeleiteten statischen Merkmale im direkten Vergleich als robuster in der Erkennungsleistung. Die betrachteten prosodischen Konturen sind der Verlauf der Grundfrequenz, der Energie, der Dauern von stimmhaften Anteilen und Pausen sowie der spektrale Verlauf. Als Vorverarbeitung erfolgt eine Wandlung mit 16bit und 11kHz sowie eine Hamming-Fensterung in quasi-stationäre Anschnitte der Länge 20ms mit 50% Überlappung. Die Grundfrequenzbestimmung selbst ist prinzipiell fehlerbehaftet, und erfolgt mit der AMDF-Methode. Diese zeichnet sich durch ihre Robustheit gegenüber Rauschen, aber auch ihrer Neigung zur Konfusion mit dominanten Formanten aus. Bei der Energie handelt es sich um die gemittelte Energie pro Fenster. Die Dauern ergeben sich aus dem Verlauf der Grundfrequenz zur Schätzung stimmhafter Anteile und der Energie bezüglich der Pausen durch Schwellwertentscheidung. Die spektralen Verläufe werden mit einer FFT errechnet.

Bei den abgeleiteten Merkmalen aus der Grundfrequenz handelt es sich im Einzelnen um:

- Mittlere Grundfrequenz, Standardabweichung
- Relativer Wert Maximum/Minimum
- Relativer Ort Maximum/Minimum
- Wert des Maximums und Mittelwert der ersten Ableitung
- Mittlere Distanz der Wendepunkte
- Standardabweichung der Distanz der Wendepunkte

Aus dem Energieverlauf ergibt sich:

- Mittlere Energie, Standardabweichung
- Relativer Wert Maximum/Minimum
- Relativer Ort Maximum/Minimum
- Mittelwert u. Median Anstiegs- und Abfallszeit der Energie
- Wert des Maximums
- Mittlere Distanz der Wendepunkte
- Standardabweichung der Distanz der Wendepunkte

Aus den Dauern werden abgeleitet:

- Rate stimmhafter Laute
- Mittlere Dauer stimmhafter Laute
- Standardabweichung stimmhafter Laute
- Mittelwert u. Median Pausenlänge

Aus dem spektralen Verlauf resultiert:

- Spektrale Energie unter 250 Hz
- Spektrale Energie unter 650 Hz

Die Merkmale werden jeweils von ihrem Mittelwert befreit und auf ihre Standardabweichung normiert. Insgesamt ergibt sich so ein 30-dimensionaler Merkmalsvektor. Die hier vorgestellten Merkmale ergaben sich aus einem Schatz an Merkmalen, der sich als besonders geeignet heraus kristallisiert hat.

4.2 Optimierung der Klassifikation

Die Wahl des Klassifikators kann hinsichtlich einer Reihe von Aspekten für den späteren Einsatz im Fahrzeug hin entscheidend sein. Zunächst scheint eine größt mögliche Erkennungsleistung wünschenswert. Diese kann durch die Fähigkeit der Lösung nichtlinearer Probleme, Diskriminativität, eigenständige Priorisierung von Merkmalen, Nachtrainierbarkeit im späteren Betrieb, eine hohe Verallgemeinerungsfähigkeit und Einsetzbarkeit diverser Feature-Sets zwischen übergeordneten Emotionsklassen positiv beeinflusst werden. Bezüglich der Effizienz sind kurze Erkennungs- und Trainingszeiten sowie ein geringer Bedarf an Referenzen wünschenswert. Aus ökonomischer Sicht ist ein geringer Speicher- und Rechenleistungsbedarf anzustreben. Als Voraussetzung für eine sinnvolle Interpretation der Ergebnisse ist die Fähigkeit eines Klassifikators absolute sowie einzelne Konfidenzen nach Emotionen bereitzustellen zu nennen. Unter Berücksichtigung dieser Ziele wurde eine Reihe unterschiedlicher Klassifikatoren eingesetzt und bezüglich ihrer maximalen Leistung bewertet. Dabei wird im Folgenden vor allem die erzielte Güte bei der Erkennung betrachtet. Die Verfahren

sind im Einzelnen: K-Nächster-Nachbar (KNN), Minimum Abstandsklassifikator mit Klassenschwerpunkt (MinAK), Gaussian Mixture Models (GMM), Multi-Layer-Perceptron (MLP) als Sonderfall neuronaler Netze sowie Support Vector Machines. Bei letzteren wurden zwei verschiedene Methoden zur Lösung von Mehrklassenproblemen untersucht (SVM und MLSVM).

Als linearer Entscheider wurde zunächst ein KNN-Klassifikator mit Euklid'schem Distanzmaß gewählt. Hierbei handelt es sich um einen Mehrheitsentscheid innerhalb der k nächsten Nachbarn. Da neue Referenzen nur gespeichert werden müssen, ergibt sich eine leichte Nachtrainierbarkeit. Allerdings ist hierdurch ein höherer Speicherbedarf erforderlich und ein zeitaufwendiger Vergleich mit allen Referenzen erforderlich. Eine Angabe über die Konfidenz ergibt sich jeweils aus den Abständen zu den Klassen oder der Trefferzahl innerhalb der ersten k Treffer.

In einem nächsten Schritt wurde daher jeweils der Klassenschwerpunkt gebildet (MinAK) um Rechenzeit und Speicherplatz zu sparen. Die Erkennungsleistung zeigte sich jedoch auf Grund der linearen Charakteristik gering.

Als statistisches Verfahren wurden daher Gaussian Mixture Models betrachtet. Es wird approximativ die tatsächliche Wahrscheinlichkeits-Dichte-Funktionen durch additive Überlagerung gewichteter Gaußverteilungen nachgebildet. Das Training wurde dabei iterativ nach dem Expectation Maximization Algorithmus vollzogen. Jede Klasse wird durch ein GM-Modell repräsentiert. Eine Entscheidung erfolgt nach der wahrscheinlichsten Klasse. Dieses Vorgehen birgt den Nachteil keiner Diskriminativität.

Um dieser Rechnung zu tragen und eine eigenständige Gewichtung der Merkmale zu erlauben, aber auch die nichtlineare Charakteristik der Emotionen im Merkmalsraum durch eine Transferfunktion zu berücksichtigen, wurde weiterhin ein mehrschichtiges Perzeptron (MLP) evaluiert. Es wurden insgesamt 30 Eingänge für die Merkmale in der Eingangsschicht eingesetzt. In einer verborgenen Schicht wurde eine variable Zahl von Neuronen mit Sigmoidfunktion betrachtet. Je Emotion wurde ein Ausgang mit Tangenshyperbolikusfunktion verwendet. Dabei wurden die Ausgänge mittels Softmax-Funktion als Konfidenzen berechnet. Als Fehlerfunktion kam die Kreuzentropie zum Einsatz und das Training wurde unter Einbezug einer Kreuz-Validierung mit Rückwärtspropagation von 10-1000 Iterationen vollzogen.

Um den Nachteil empirischer Risiko-Minimierung und der damit verbundenen Gefahr einer Überadaption eines neuronalen Netzes zu kompensieren, wurden schließlich Support Vector Machines (SVM) als weiteres Verfahren in die Betrachtung mit einbezogen. Sie erreichen eine hohe Generalisierungsfähigkeit durch strukturelle Risiko-Minimierung mittels einer Fehler-Obergrenze. Die Behandlung nichtlinearer Probleme wird bei dem prinzipiell linearen Klassifikator durch Abbildung in einen im Allgemeinen höherdimensionalen Raum erzielt. Hierbei kommt eine dem Problem angepasste Mapping-Funktion zum Einsatz. Diskriminativität wird durch eine optimale Trennebene zwischen zwei Klassen erzielt. Die Trennebene wird durch Stützvektoren beschrieben, was eine Reduktion der Daten erlaubt. Es existiert eine Reihe unterschiedlicher Ansätze um mehrere Klassen mit SVMs trennen zu können. In der vorgestellten Arbeit wurden zwei Verfahren betrachtet: Ein Training jeder Klasse gegen die jeweils anderen Klassen mit einer Maximum-Likelihood-Entscheidung (SVM) und ein Ansatz in mehreren Ebenen, Multi-Layer-Support Vector Machines (MLSVM) genannt. Die Entscheidung erfolgt dabei jeweils ebenen- und paarweise wie in der folgenden Abbildung zu sehen:

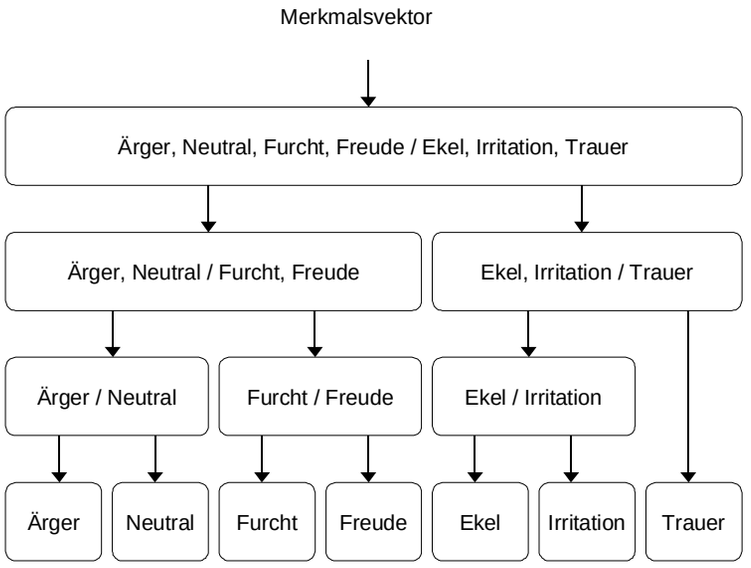

Abbildung 2: Beispiel Anordnung der Emotionen bei MLSVMs

Hierdurch lassen sich semantische Zusammenhänge in Form von Expertenwissen nach der Regel „schwer trennbare zusammenfassen“ modellieren. Das heißt die Anordnung der Klassen auf die Ebenen beeinflusst im entscheidenden Maße die Erkennungsgüte. Dies lässt sich auch durch eine vorab Evaluation mit SVMs automatisieren. Aus den Konfusionsmatrizen der Emotionen erfolgt die Zuordnung auf die

Ebenen und Paare. Die insgesamt höhere Erkennungsrate der MLSVMs lässt sich nur unter Verlust der Einzelkonfidenzen erzielen.

Um die Vorteile verschiedener genannter Klassifikatoren zu vereinen, wurde außerdem eine Reihe hybrider Ansätze evaluiert. Diese sind im Einzelnen ein GMM, welches ein MLP speist, SVMs die ein MLP speisen und ein MLP, welches SVMs speist. Dabei wurden jeweils die Ausgänge der SVMs oder GMMs an ein neuronales Netz übergeben, beziehungsweise im letzten genannten Fall der Ausgang eines MLPs an SVMs übergeben.

5. Integration semantischer und prosodischer Aspekte

Um die Ergebnisse der semantischen und prosodischen Analyse zu integrieren bieten sich grundsätzlich drei Verfahren an: eine frühe Vereinigung auf Merkmalsebene, Early Feature Fusion genannt, eine späte Vereinigung nach der Entscheidung, Late Semantic Fusion genannt, sowie die sogenannte Soft Decision Fusion. Während die Early Feature Fusion auf Grund des divergenten zeitlichen Verhaltens hier nur bedingt umsetzbar ist, ist die Late Semantic Fusion mit einem Wissensverlust zu Gunsten einer modularen Natur verbunden. Der Wissensverlust soll in der Soft Decision Fusion durch weiche Entscheidungen unter Sicht der Konfidenzen konkurrierender Hypothesen umgangen werden, was einen sinnvollen Kompromiss erlaubt. Sollen nur die Ergebnisse prosodischer und semantischer Analyse betrachtet werden, eignet sich etwa ein neuronales Netz für die Integration. Als Eingang sind hier jeweils die Konfidenzen jeder Instanz über jede Emotion zu verwenden. Das Netz kann selbstständig lernen, welcher Instanz bezüglich welcher Emotion besonders zu vertrauen ist, und in welcher Weise mehrdeutige Szenarien zu entscheiden sind. Kommt eine der beiden Instanzen zu keiner Entscheidung, kann direkt das Ergebnis der jeweils anderen Instanz verwendet werden. Liegt kein ausreichendes Trainingsmaterial über beide Ansätze vor, oder sollen weitere Modalitäten oder Kontextwissen in die Analyse mit einfließen, können alternativ Bayes'sche Netze unter Ausnutzung der beschriebenen Vorteile eingesetzt werden.

6. Einsatz im Fahrzeug

Im Gegensatz zur Spracherkennung im Fahrzeug erscheint es bei der sprachlichen Emotionserkennung wenig sinnvoll einen Push-to-Talk-Button oder ähnliche benutzerinitiierte Segmentierungen einzusetzen. Dies beruht auf der Tatsache, dass der Benutzer die Emotionen in der Regel nicht explizit dem System mitteilen wird, und verlangt eine robuste Erkennung im kontinuierlichen Signalverlauf. Um im offenen Mikrofonbetrieb Sprache von Störgeräusch zu unterscheiden, wurde eine

Diskrimination auf Basis von 256 FFT Merkmalen unter Verwendung von SVMs mit Radial-Basis-Kernel (RBF) eingesetzt. Diese wurde mit Sprachsamples und Störgeräuschen aus der Fahrzeugumgebung trainiert. In unseren Tests ergaben sich hierbei 99.1% Erkennungsleistung. Um weiterhin fremdsprachliche Äußerungen etwa aus dem Radio, Telefon oder von Mitfahrern filtern zu können, wird eine schritthaltende Sprecherverifizierung verwendet. Hierzu setzen wir 17 MFCC-Koeffizienten und ihre Ableitungen ebenfalls auf Basis von RBF-SVMs ein. Dabei werden Angreifer-Modelle fremder Sprecher in einzelnen SVMs jeweils gegen den Sprecher trainiert, oder der Abstand zur Trennebene als Maß für die Sicherheit eingesetzt. Als besondere Herausforderung sind hier die kurzen emotionalen Äußerungen mit 3,1 Sekunden im Mittel zu vergleichsweise wünschenswerten 30 Sekunden in üblichen Sprecherverifikationsapplikationen mit starken emotionsbedingten Verzerrungen zu nennen. Bei einer Datenbank mit 18 typischen Sprechermodellen ergab sich eine Erkennungsrate von 96.7%. Gemeinsam lassen sich die beiden Modelle als MLSVM darstellen:

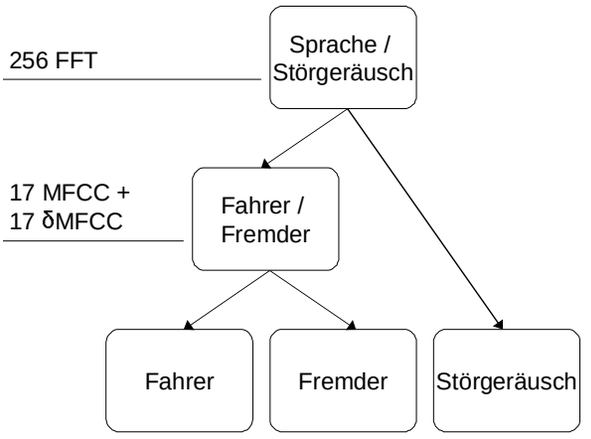

Abbildung 3: MLSVM für die kontinuierliche Emotionserkennung

Als weitere Sicherheit wird eine Mindest-Konfidenz über eine geschätzte Emotion vorausgesetzt, ehe auf diese reagiert wird.

7. Resultate und Diskussion

In diesem Abschnitt sollen die Ergebnisse, die mit den vorgestellten Methoden erzielten wurden, präsentiert werden.

7.1 Vergleichsbasen

Um eine Einschätzung über das Urteilsvermögen des Menschen bezüglich der Emotion zu ermöglichen, wurde in einer Studie vorab der vorgestellte Sprachkorpus von zehn Testpersonen zugeordnet. Dabei musste jeder Proband sowohl sich selbst, als auch eine Stichprobe

mehrer anderer Sprecher klassifizieren. Die mittlere Erkennungsleistung bei der Selbstklassifikation beläuft sich dabei auf 85,5%. Bei der Fremdklassifizierung sinkt diese Leistung auf 64,7% bei jeweils sieben Emotionen nach dem MPEG4-Standard plus einen neutralen Zustand.

7.2 Ergebnisse der Evaluation

Zunächst soll die Erkennungsleistung bei der akustischen Analyse vorgestellt werden. Hierzu wurde jeweils dreimal mit zwei Dritteln der Daten trainiert und einem Drittel evaluiert. Die mittlere optimale Erkennungsleistung bei sieben Emotionen mit jeweils 700 Samples je Sprecher lag bei 89,3%. Wurde mehr als ein Sprecher gleichzeitig trainiert und evaluiert, sank diese Leistung im Mittel auf 76,3%. Im Folgenden sind die maximal erzielten Erkennungsleistungen nach Verfahren unter Angabe der optimalen Parameter aufgeschlüsselt:

- MinAK: 66,8%
- KNN: 75,5% bei k=1
- GMM: 83,6% bei 8 Mixturen
- MLP: 84,5% bei 100 Neuronen im Hidden Layer
- SVM: 86,3% bei polynomial Kernel
- MLSVM: 89,3%
- GMM->MLP: 71,8% bei 8 Mixturen und 1200 Neuronen
- SVM->MLP: 85,7% bei 400 Neuronen
- MLP->SVM: 87,3% bei RBF-Kernel und 100 Neuronen

Mit den betrachteten Sprechern ergaben sich dabei folgende Konfusionen:

Emotion	Ärger	Ekel	Furcht	Irritation	Freude	Neutral	Trauer
Ärger	**0.716**	0.010	0.069	0.010	0.049	0.098	0.049
Ekel	0.099	**0.733**	0.059	0.000	0.059	0.050	0.000
Furcht	0.000	0.040	**0.800**	0.080	0.000	0.000	0.080
Irritation	0.000	0.020	0.049	**0.824**	0.039	0.010	0.059
Freude	0.000	0.040	0.000	0.069	**0.842**	0.000	0.050
Neutral	0.029	0.039	0.000	0.010	0.000	**0.863**	0.059
Trauer	0.040	0.139	0.040	0.010	0.030	0.079	**0.663**

Abbildung 4: Konfusion der Emotionen

Wie zu sehen ist, lassen sich die Emotionen mit unterschiedlicher Güte erkennen.

Die Emotionserkennungsrate bei der Phrasenanalyse liegt über 95%, hängt jedoch stark von der Erkennungsleistung des Spracherkenners ab.

Durch die Integration beider Modelle konnte ein klarer Zugewinn erzielt werden. Dabei erhöhte sich neben der Erkennungsleistung auch die Gesamtzahl an Äußerungen, in denen eine emotionale Schätzung erfolgen konnte. Als weiteres Ergebnis zeichnete sich eine unterschiedliche Eignung dieser Ansätze nach Emotionen ab. So konnte etwa Irritation und der neutrale Zustand besonders gut durch die Akustik festgestellt werden, wohingegen eine Unterscheidung zwischen Freude und Ärger besonders mit dem semantischen Ansatz gelang.

7.3 Fazit und Ausblick

In der akustischen Analyse konnte sich besonders der Einsatz von MLSVMs als vorteilhaft erweisen. Hier gelang eine Erkennungsrate bei Training mit dem späteren Sprecher in der Güte eines menschlichen Entscheiders. Bei Evaluation mit trainingsfremden Sprechern ergab sich eine um 6% schlechtere Leistung als bei einem menschlichen Entscheider. Dies zeigt die starke Sprecherabhängigkeit des verwendeten Ansatzes. Hier kann eventuell eine deutlich größere Datenbank mit mehr Sprechern im Training abhelfen. Besonders in der sprecher-unabhängigen Erkennung erwies sich der phrasenbasierte Ansatz als robuster.

Die erkannten Emotionen konnten in einem Demonstrator erfolgreich in die Kommunikation im Fahrzeug integriert werden.

Als weitere Ziele bleiben Langzeituntersuchungen im Feld mit mehr Probanden und spontaneren Emotionen. Um eine höhere Unabhängigkeit vom Sprecher zu erreichen sollen emotionale Typmodelle und Strategien zur Benutzeradaption untersucht werden. Des Weiteren können ergänzende Merkmale in den Analyseprozess integriert oder weitere emotionale Zustände wie Schmerz bei einem Unfall betrachtet werden.

Danksagung

Die in diesem Beitrag vorgestellte Arbeit wurde zu großen Teilen vom Projekt FERMUS, einer Zusammenarbeit der Partner BMW Group, DaimlerChrysler, SiemensVDO und dem Lehrstuhl für Mensch-Maschine-Kommunikation an der Technischen Universität München unterstützt. Das Projekt steht für fehlerrobuste multimodale Sprachdialoge im Fahrzeug.

Literatur

[03dev1] Devillers, L.; Lamel, L.: *Emotion Detection in Task-Oriented Dialogs*. Tagungsband ICME 2003, 06.-09.07.2003, Baltimore, MD, USA. IEEE. Multimedia Human-Machine Interface and Interaction I, Vol. III, S. 549-552. Auch CD-ROM.

[01lee1] C. M. Lee, R. Pieraccini, “Combining acoustic and language information for emotion recognition,” ICSLP 2002, USA, 2002

[02mcg1] McGlaun, G.; Althoff, F.; Schuller, B.; Lang, M.: *A new technique for adjusting distraction moments in multi-tasking non-field usability tests.* Tagungsband Intern. Conference on Human Factors in Computing Systems CHI 02, Minneapolis, USA, 20.-25.04.2002. Hrsg.: Terveen; Wixon; Comstock; Sasse. ACM SIGCHI, New York, 2002. S. 666-667.

[02nie1] Nieschulz, R.; Schuller, B., Geiger, M.; Neuss, R.: *Aspekte effizienten Usability Engineerings.* Themenheft der Zeitschrift "it+ti", Schwerpunktthema "Usability Engeneering", Oldenbourg Wissenschaftsverlag, München, 1/2002, S. 23-30.

[02sch1] Schuller, B.; Lang, M.; Rigoll, G.: *Automatic Emotion Recognition by the Speech Signal.* Tagungsband SCI 2002, 6th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, Florida, USA, 14.-18.07.2002. Proceedings Vol. IX, "Image, Acoustic, Speech and Signal Processing II", Hrsg.: Callaos, IIIS, S. 367-372.

[02sch2] Schuller, B.; Lang, M.; Rigoll, G.: *Multimodal Emotion Recognition in Audiovisual Communication.* CD-ROM-Proceedings IEEE International Conference on Multimedia and Expo, ICME 2002, Lausanne, Schweiz, 26.-29.08.2002. Ed.: SuvoSoft Oy Ltd., Tampere, Finnland.

[02sch3] Schuller, B.: *Towards intuitive speech interaction by the integration of emotional aspects.* CD-ROM-Proceedings IEEE International Conference on Systems, Man and Cybernetics, SMC 2002, "Bridging the Digital Divide", Yasmine Hammamet, Tunesien, 6.-9.10.2002. Eds.: A. El Kamel, K. Mellouli, P. Borne. Vol. 6, WA2N1.

[02sch4] Schuller, B.; Althoff, F.; McGlaun, G.; Lang, M.; Rigoll, G.: *Towards Automation of Usability Studies.* CD-ROM-Proceedings IEEE International Conference on Systems, Man and Cybernetics, SMC 2002, "Bridging the Digital Divide", Yasmine Hammamet, Tunesien, 6.-9.10.2002. Eds.: A. El Kamel, K. Mellouli, P. Borne. Vol. 4, TP1N6.

[03sch1] Schuller, B.; Rigoll, G., Lang, M.: *Hidden Markov Model-Based Speech Emotion Recognition.* Tagungsband ICASSP 2003, 06.04.-10.04.2003, Hong Kong, China. IEEE 2003. Vol. II, S. 1-4. Auch CD-ROM.