Schlussbericht

zum Teilvorhaben

Sprachtechnologien für sprachliche Interaktion

im Verbundprojekt

Kooperative Fahrer-Fahrzeug-Interaktion: Sichere und effiziente Interaktion mit autonomen Fahrzeugen

Förderkennzeichen: 16SV7627

Teil I: Kurzbericht

Dr.-Ing. Volker Fischer
EML European Media Laboratory GmbH
Berliner Straße 45
D-69120 Heidelberg

E-Mail: volker.fischer@eml.org
Tel.: +49 6221 679 26 26

Kurzbericht

Aufgabenstellung

Das Gesamtziel des Verbundprojektes „Kooperative Fahrer-Fahrzeug-Interaktion: Sichere und effiziente Interaktion mit autonomen Fahrzeugen“ (kurz: „**KoFFI**“) bestand in der Entwicklung eines ganzheitlichen, psychologisch fundierten Konzepts zur kooperativen Fahrer-Fahrzeug-Interaktion für hochautomatisierte Fahrzeuge. Die Hauptaufgabe des von EML bearbeiteten Teilvorhabens „Sprachtechnologien für sprachliche Interaktion“ bestand in der Entwicklung, Anpassung und Evaluierung von Methoden der automatischen Spracherkennung, sowie in deren Integration in das vom Projektpartner Daimler AG entwickelte Sprachdialog-System. Ziel war es, durch eine hochakkurate Spracherkennung und die Untersuchung des Sprachstils eine situations- und nutzergerechte sprachliche Interaktion beim (teil-)automatisierten Fahren zu ermöglichen.

Wissenschaftlicher und technischer Stand zu Projektbeginn

Durch die Ablösung von „klassischen“ Hidden-Markov-Modellen mit Gauß’schen Emissionswahrscheinlichkeiten (HMM/GMM) durch tiefe neuronale Netzwerke hat die automatische Sprach*erkennung* in den vergangenen 5 – 10 Jahren eine dramatische Leistungssteigerung erfahren [1, 2]. Während zu Projektbeginn im Bereich der akustischen Modellierung die Verwendung von tiefen Modellen mit einfacher Vorwärtsarchitektur (*feed forward neural networks, FFNN*) dominierte, wird der Stand der Technik nunmehr durch den Einsatz rekurrenter neuronaler Netzwerke (*RNNs*) definiert; als wichtigste Ausprägungen können LSTMs (*long short term memory*, [3, 4]) und GRUs (*gated recurrent units*, [5]) angesehen werden. Es ist davon auszugehen, dass die meisten der im Automobil eingesetzten kommerziellen Spracherkenner mittlerweile LSTMs oder GRUs zur akustischen Modellierung einsetzen.

Eine vergleichbare Entwicklung ist im Bereich der statistischen Sprachmodellierung festzustellen. Nach der Einführung leistungsstarker Methoden zur numerischen Repräsentation von Wörtern bzw. deren Semantik [6, 7] wurden konventionelle n-gram Sprachmodelle durch FFNNs und RNNs *ergänzt* (z.B. [8, 9, 10]). Die gängige Strategie zur Integration NN-basierter Sprachmodelle in die Suchstrategie eines Spracherkenners besteht bislang darin, Hypothesen zunächst mit einem konventionellen n-gram Sprachmodell zu erzeugen und NN-basierte Modelle lediglich *offline* zum *Rescoring* einzusetzen. Ansätze zur effizienten Nutzung RNN-basierter Sprachmodelle in einem *Online*-Spracherkenner finden sich erst in neueren Arbeiten [11].

Die durchgeführten Untersuchungen zur *textbasierten* Sprachstilklassifikation berühren das Gebiet der Verarbeitung natürlicher Sprache (*Natural Language Processing, NLP*). Einen Überblick über zu Projektbeginn dominierende *Machine-Learning*-Verfahren gibt [12]. Mit dem Einzug großer Korpora und leistungsfähiger Hardware stellen neuronale Ansätze mittlerweile jedoch auch im NLP-Bereich für eine Vielzahl von Aufgabenstellungen den Stand der Technik dar [13]. Zu erwähnen sind insbesondere sogenannte Transformer [14, 15], die sehr exakte semantischen Repräsentation (der Wörter eines Korpus) darstellen bzw. erzeugen. Vortrainierte Transformer-Netzwerke, die mit relativ wenigen Daten an die jeweilige Domäne und Klassifikationsaufgabe adaptiert werden können, sind nach unserem Kenntnisstand bislang jedoch nur für das Englische verfügbar.

Ablauf des Vorhabens

Am Verbundvorhaben KoFFI waren sechs Partner beteiligt, nämlich:

- Robert Bosch GmbH
- Daimler AG
- EML European Media Laboratory GmbH
- Hochschule Heilbronn (UniTy Lab)
- Hochschule der Medien Stuttgart (HdM, Institut für Digitale Ethik)
- Universität Ulm (Institut für Psychologie und Institut für Medieninformatik)

Verbundkoordinator war die Robert Bosch GmbH, Abteilung Car Multimedia. Die Kooperationsvereinbarung wurde von EML am 14.10.2016 geschlossen; die Projektlaufzeit war vom 01.11.2016 bis zum 31.10.2019. Die gesamte Fördersumme betrug ca. 3.6 Millionen Euro, davon entfielen 286.816 Euro auf EML.

Seitens EML wurden über die gesamte Projektlaufzeit Mitarbeiter mit großer Erfahrung im Bereich von Sprachtechnologien sowie Software-Entwickler mit langjähriger Praxis im Bereich der Architektur komplexer Software-Systeme und der Spezifikation von Schnittstellen eingesetzt. Nach dem Ausscheiden von Dr. Siegfried Kunzmann ging die Projektleitung zum 01.01. 2017 auf Dr. Volker Fischer über.

Das KoFFI-Projekt war in acht Arbeitspakete (APs) aufgeteilt. EML war an allen Arbeitspaketen beteiligt; der Schwerpunkt der Arbeiten lag jedoch auf dem Arbeitspaket zur sprachlichen Interaktion (AP5), gefolgt von AP3 und AP7, deren Themen die Definition von Use-Cases für die Gestaltung der kooperativen Mensch-Maschine-Schnittstelle und die Integration des Gesamtsystems waren.

Regelmäßige, etwa alle 3 Monate stattfindende Workshops wurden zur Koordination sämtlicher Aktivitäten unter den Verbundpartnern, zur Arbeit an den Use-Cases (AP3), und zum Austausch von (Zwischen-)Ergebnissen genutzt. Mit dem in AP5 federführend beteiligten Projektpartner Daimler AG wurde bei der Definition der Anforderungen an die situationsbezogene Spracherkennung, deren Realisierung, sowie der experimentellen Evaluierung eng und unkompliziert zusammengearbeitet. Am EML wurden alle Arbeiten planungsgemäß durchgeführt und alle Meilensteine erreicht.

Während der gesamten Laufzeit hat EML auch starken Wert auf die Weiterverbreitung der im Projekt erzielten Ergebnisse gelegt. Dazu wurden ausgewählte Komponenten des entwickelten Spracherkenners auf internationalen und nationalen Tagungen vorgestellt, reguläre Messeauftritte von EML zum Anlass für Pressemitteilungen über KoFFI genommen und Besucher und Gäste von EML – etwa aus Anlass des jährlich veranstalteten Heidelberg Laureate Foundation Forums – über das Projekt informiert. Ein Highlight der Öffentlichkeitsarbeit war sicherlich der Besuch von Bundesministerin Karliczek auf dem gemeinsam mit allen Verbundpartnern betreuten Stand auf der CEBIT 2018.

Wesentliche Ergebnisse

Zur *automatischen Spracherkennung* konnte auf existierende Sprachdatensammlungen für das Deutsche – insbesondere auch auf die im Automobil gesammelten SpeechDat-CAR Daten – zugegriffen werden. Das dort vorhandene Textmaterial deckt auch Kommandos für die Steuerung von Sekundärfunktionen und Infotainment-Systemen im Fahrzeug ab. Andere, im Kontext des *automatisierten* Fahrens benötigte Sätze oder Phrasen sind in keiner verfügbaren Datensammlung vorhanden; sie wurden von EML und Daimler in begrenztem Umfang gesammelt und von EML zur Anpassung des statistischen Sprachmodells an die KoFFI-Domäne bzw. Use-Cases benutzt. Die Erkennungsleistung des realisierten Spracherkennungssystems gestattet es dem Dialogsystem mehr als 95 Prozent aller Anfragen korrekt zu erledigen.

Im Bereich der (textbasierten) *Sprachstilklassifikation* wurden Studien zur Bestimmung der Abweichung vom „normalen“ Sprachstil, zur Erkennung der Fahrsituation und zur Priorisierung von Äußerungen durch die Wortwahl des Benutzers durchgeführt. Weit mehr als im Bereich der Spracherkennung waren diese Arbeiten vom Fehlen geeigneter und in der Forschung etablierter Korpora aus der Domäne gekennzeichnet. Zur Entwicklung der NN-basierten Klassifikatoren und zur Fortschrittskontrolle hat EML daher vorrangig domänenfremde, englischsprachige Korpora eingesetzt und die erzielten Ergebnisse abschließend an den in geringem Umfang im Projekt gesammelten Daten verifiziert. In den Untersuchungen konnte die in der Fachliteratur festgestellte Überlegenheit neuronale Ansätze bestätigt werden.

Zusammenarbeit mit anderen Stellen

Bei den Untersuchungen zur Klassifikation von Benutzeräußerungen arbeitete EML eng mit dem Projektpartner Daimler AG zusammen. Der von allen Projektpartnern gemeinsam realisierte Demonstrator nutzt die von EML spezifizierte Schnittstelle zur *EML Transcription Platform,* einer hochskalierbaren Umgebung für eine server-basierte Spracherkennung. An den Untersuchungen zu ethischen, rechtlichen und sozialen Implikationen (ELSI) war EML insbesondere im Bereich der Datensicherheit und des „Privacy by Design“ für das Sprachdialogsystem beteiligt.

Schlussbericht

zum Teilvorhaben

Sprachtechnologien für sprachliche Interaktion

im Verbundprojekt

Kooperative Fahrer-Fahrzeug-Interaktion: Sichere und effiziente Interaktion mit autonomen Fahrzeugen

Förderkennzeichen: 16SV7627

Teil II: Eingehende Darstellung

Dr.-Ing. Volker Fischer
EML European Media Laboratory GmbH
Berliner Straße 45
D-69120 Heidelberg

E-Mail: volker.fischer@eml.org
Tel.: +49 6221 679 26 26

Eingehende Darstellung

Erzielte Ergebnisse in Gegenüberstellung zur ursprünglichen Planung

Übersicht. Das Teilvorhaben „Sprachtechnologien für sprachliche Interaktion“ war an allen acht Arbeitspaketen des Verbundprojekts beteiligt:

- In AP1 (Projektmanagement) wurde das Erreichen der Projektziele und Meilensteine und die Einhaltung der Kosten sichergestellt.
- In AP2 (ELSI) wurden die Projektpartner über die eingesetzten Spracherkennungstechnologien informiert und das eigene Vorgehen auf die Einhaltung der von HdM gegebenen Empfehlungen (und der DSGVO) überprüft.
- In AP3 (Anforderungen an KoFFI) wurde zur Definition der Anforderungen an die sprachliche Interaktion beigetragen.
- An AP4 (Grafisch-haptische Interaktion) und AP6 (Fahrer-Fahrzeug-Modellierung) war das Teilvorhaben unterstützend aus Sicht der sprachlichen Interaktion beteiligt.
- In AP5 (Sprachliche Interaktion) wurden sämtliche der im Folgenden ausführlich beschriebenen Arbeiten zur akustischen Modellierung, zur Domänenmodellierung, und zur Sprachstilklassifikation durchgeführt, sowie die notwendigen Änderungen am Spracherkenner bzw. der *EML Transcription Platform* vorgenommen.
- In AP7 (Integration des Gesamtsystems) wurde an der SW-Integration und der Anbindung der Sprachtechnologien an das Dialogsystem bzw. den Demonstrator gearbeitet.
- In AP8 (Demonstration und Validierung) wurden zahlreiche Experimente zur Einstellung der Laufzeitparameter durchgeführt und verschiedene neuronale Architekturen zur Textklassifikation evaluiert.

Die Kernaufgabe von EML war die Untersuchung und Bereitstellung von Spracherkennungstechnologien für die nutzer- und situationsgerechte sprachliche Interaktion beim (semi-)autonomen Fahren sowie deren Integration in das von der Daimler AG entwickelte Sprachdialog-System und den Demonstrator (APs 5, 7, und 8). Im Fokus der Arbeiten stand dabei die Weiter- bzw. Neuentwicklung von Modellen und Algorithmen für den bislang vorrangig in Call-Center-Anwendungen eingesetzten EML-Spracherkenner; Ziel der Leistungsoptimierung für das Szenario des (semi-)autonomen Fahrens war es, eine *sichere*, *zuverlässige*, und *komfortable* sprach-basierte Interaktion zwischen Fahrer und Fahrzeug zu ermöglichen. Der Aspekt der *situationsgerechten* Interaktion wurde durch ergänzende Studien zur Sprachstilklassifikation adressiert, die sich mit der Bestimmung der Fahrsituation (Übernahme, Übergabe, usw.), der Ermittlung von Abweichungen vom „normalen“ Sprachstil, sowie der Bestimmung der Priorität einer Äußerung befassten.

Akustisch-linguistische Modellierung. Im Bereich der akustischen Modellierung für die automatischen Spracherkennung hat EML während der Laufzeit des Projektes den Übergang von GMM/HMM basierten TANDEM-Modellen hin zu hybriden NN- und (bidirektionalen) LSTM-basierten Modellen vollzogen [16, 17, 18]. Die dazu notwendigen Arbeiten stützen sich insbesondere auf [4, 19, 20] und setzen – neben der Verwendung der oben erwähnten SpeechDat-CAR Daten [21] – auch Methoden der „künstlichen Verschmutzung“ weiteren Trainingsmaterials zur Anpassung an das akustische Umfeld im Fahrzeug [22] sowie zur Simulation einer größeren Sprechervielfalt [23, 24] ein.

Der für das KoFFI-Szenario charakteristische Aspekt der *sicherheitsrelevanten Sprachinteraktion* wurde durch die gleichzeitige Optimierung von Erkennungsgenauigkeit und -geschwindigkeit des Spracherkenners adressiert. Dazu wurden Methoden zur Elimination von Netzwerkparametern [25] von Vorwärtsnetzwerken auf bidirektionale LSTMs übertragen, eine *online*-fähige BLSTM-Variante entwickelt [18], und eine robuste, online-fähige Komponente zur Sprachdetektion (*voice activity detection*) mit kurzem Entscheidungshorizont entworfen und integriert [26].

Die Arbeiten zur akustisch-linguistischen Modellierung wurden durch die Anpassung von n-gram Sprachmodell und Vokabular an die in AP3 definierten Use-cases komplettiert. Basis für die Adaption waren dabei ein am EML verfügbares, allgemeines Sprachmodell für das Deutsche mit etwa 1.5 Millionen Vollformen sowie eine EML-eigene Toolbox [27]. Als Datenmaterial diente vorrangig der im Projekt gesammelte Textkorpus mit

Äußerungen aus verschiedenen Fahrsituationen (siehe unten); ergänzend wurden auch wenige Wörter – insbesondere Imperative aus dem Problemkreis („überhole“, „übernehme“) oder Eigennamen („KoFFI“, „Mercedes“) – zum Lexikon hinzugefügt.

Zur Evaluierung wurden zwei Testmengen verwendet:

- T1 lag zu Projektbeginn am EML vor und umfasst 1506 Audiodateien mit einer Gesamtlänge von 77.5 Minuten (einschl. Stille) und 7441 Referenzwörtern. Anwendungsgebiet ist die Interaktion mit dem Infotainment-System des Fahrzeugs sowie die Steuerung von Sekundär- oder Komfortfunktionen (Licht, Klimaanlage, usw.).
- T2 wurde vom Projektpartner Daimler AG während der Projektlaufzeit gesammelt und umfasst 1608 Audiodateien mit einer Gesamtlänge von 55.5 Minuten und 7446 Referenzwörtern. Etwaige Abweichungen der Sprecher(innen) vom Wortlaut des vorgegebenen Skripts wurden durch eine Mitarbeiterin von EML korrigiert. Das Skript enthält typische Kommandos aus dem Szenario des (semi-)autonomen Fahrens, also beispielsweise Anweisungen zur Übernahme und Übergabe („Ab jetzt fahre ich“, „KoFFI, fahre mich zur Arbeit“) oder auch Fragen zum Systemverhalten („KoFFI, warum überholst Du nicht?“).

Tabelle 1 zeigt den im Projekt erzielten Fortschritt — gemessen als Wortfehlerrate — beim Übergang zu (bidirektionalen) LSTMs, die mit den für das automobile Szenario aufbereiteten Daten trainiert wurden. Die Ergebnisse für T2 belegen auch die Notwendigkeit der Anpassung des Sprachmodells an das KoFFI-Szenario, da letzteres im allgemeinen Sprachmodell (allg. LM) offensichtlich nur schwach abgedeckt ist. Mit der für T2 erzielten Wortfehlerrate von 8.3 Prozent ist das Dialogsystem in der Lage mehr als 95 Prozent aller Anfragen korrekt zu behandeln.

	T1 (Infotainment)		**T2 (autonomes Fahren)**	
	allg. LM	KoFFI LM	allg. LM	KoFFI LM
Basis AM	41.6	NA	40.0	9.3
BLSTM	23.6	22.6	25.5	8.3

Tabelle 1: Wortfehlerraten (in Prozent) für zwei verschiedene Testmengen vor und nach Übergang zur LSTM-basierten akustischen Modellierung (AM), sowie mit und ohne Anpassung des Sprachmodells (LM).

Spracherkenner und Plattform. Die *EML Transcription Platform* [28, 29] ist eine hochskalierbare Softwareumgebung zum server-basierten Betrieb des EML Spracherkenners, der eine für den kommerziellen Betrieb erweiterte und optimierte Variante des in [30, 31] beschriebenen Spracherkenners mit dynamischer Zustandsraumsuche ist. Zur Nutzung durch das in KoFFI verwendete Sprachdialog-System wurden sowohl der Spracherkenner als auch die Plattform angepasst.

Neben der bereits erwähnten Erweiterung des Spracherkenners um eine robuste online-fähige Sprachdetektion wurde auch eine Möglichkeit zur parallelen Erkennung mehrerer Grammatiken und/oder Sprachmodelle geschaffen. Im KoFFI-Szenario soll damit einerseits die Notwendigkeit zur Aktivierung des Spracherkenners per Tastendruck durch die Fahrerin *(„push-to-talk“*) entfallen; andererseits soll eine situationsbezogene Spracherkennung besser unterstützt werden. Jede Grammatik bekommt dabei einen eigenen Suchraum zugeordnet, dessen Worthypothesen in einem separaten Thread dynamisch erweitert bzw. durchsucht werden. Die situationsbezogene Aktivierung von Grammatiken und die Auswahl von Ergebnissen bleibt dabei Aufgabe der Anwendung, im KoFFI -Szenario also des Sprachdialog-Systems.

Innerhalb der *EML Transcription Platform* sorgt der Streaming Server für die bidirektionale und synchrone Kommunikation zwischen Spracherkenner und Anwendungsprogramm. Für das KoFFI-Szenario wurden die Ergebnisse – erkannte Wörter, Aussprachen, zugehörige Zeitmarken und Konfidenzen – um einige (statistische) Kenngrößen angereichert, die vom Sprachdialog-System verwendet werden können. Zu nennen sind hier die auf die jeweilige Äußerung bezogene Sprechgeschwindigkeit und der Anteil an Sprechpausen, durch

erkannte Konjunktionen angezeigte Satz- oder Phrasengrenzen, und ein Indikator für die Dringlichkeit einer Äußerung (siehe unten).

Sprachstilklassifikation. Während die im beschriebenen Modellierungstechniken und Funktionserweiterungen mittlerweile in nahezu allen von EML unterstützten Sprachen und Anwendungen genutzt werden, waren die Untersuchungen zur *textbasierten* Sprachstilklassifikation stark auf das KoFFI-Szenario ausgerichtet und konzentrierten sich auf drei Aspekte, nämlich:

- die Erkennung von Abweichungen von einem – in einer gegebenen Fahrsituation – „normalen Sprachstil“,
- die Bestimmung der Fahrsituation selbst, sowie
- die Bestimmung der Dringlichkeit einer Nutzeräußerung.

Das bereits in Abschnitt 1.3 erwähnte Fehlen von für die Domäne geeigneten Textkorpora zum Training von Klassifikatoren erschwerte das geplante Vorgehen erheblich. Auch konnte die in der Vorhabensbeschreibung anvisierte Datensammlung nicht durchgeführt werden, da weder genügend freiwillige Teilnehmer noch geschulte, mit dem Szenario vertraute Hilfskräfte zur Annotation gewonnen werden konnten. Die Entwicklung einer Trainingsumgebung erfolgte daher zunächst mit frei verfügbaren, domänenfremden englisch-sprachigen Korpora [32, 33]. Später wurden zwei vom Projektpartner Daimler AG unter Beteiligung von EML erstellte Korpora zur Evaluierung verschiedener Algorithmen herangezogen:

- K1 ist ein Textkorpus mit typischen Nutzeräußerungen aus 6 verschiedenen Fahrsituationen bzw. Befehlsgruppen (Fahrtbeginn, Übergabe, Übernahme, Überholmanöver, Infotainment, Zustimmung) und umfasst 4624 Wörter, davon 713 verschiedene. An der Datensammlung beteiligten sich 94 Personen, von denen auch Alter und Geschlecht festgehalten wurden. Außerdem bewerteten die Probanden ihren Grad der Erfahrung mit Sprachdialog-System und gaben die Häufigkeit der Nutzung solcher Systeme an. Eine Handklassifikation der insgesamt 664 Äußerungen als „normal“ bzw. „abweichend“ ist nicht vorhanden.
- K2 ist ein Audiokorpus mit 10400 Aufnahmen von 40 Probanden, der durch einen *Gamification*-Ansatz gewonnen wurde, bei dem die Teilnehmer in einem sprach-gesteuerten Computerspiel in unterschiedlich dringende (Spiel-)Situationen versetzt wurden. Der Ansatz wurde gewählt, da ein Fahrsimulator nicht zur Verfügung stand und sich ein „reales“ Vorgehen von selbst verbietet. Es werden maximal 4 Dringlichkeitsstufen unterschieden.

Mit Hilfe von K1 wurde zunächst ein empiristischer Ansatz (das sog. *Latent Semantic Mapping (LSM)* [34]) zur Textklassifikation untersucht. Das auf der Singulärwertzerlegung von Word-Dokument-Matrizen beruhende Verfahren lieferte eine Fehlerrate von 12.1 Prozent für die Unterscheidung der 6 Fahrsituationen. Durch die Verwendung von N-gram Merkmalen [35] und einem einfaches DNN mit 2 verborgenen Schichten wurde eine Fehlerrate von 5.8 Prozent erreicht, und mit Hilfe eines CNN (*convolutional neural network*) konnte schließlich eine Fehlerrate von 1.2 Prozent erzielt werden. Damit wurde der aus der Literatur bekannte *state-of-the-art* nachvollzogen; die Ergebnisse sollten jedoch aufgrund der äußerst geringen Datenmengen nicht überbewertet werden.

Zur Bestimmung der Abweichungen von einer für die jeweilige Situation normalen Wortwahl wurden die vom LSM erzeugten semantischen Repräsentationen der Äußerungen geclustert und Abstandsmessungen durchgeführt. Ein großer Abstand vom Zentrum eines Clusters sollte demnach einer starken Abweichung von der normalen Wortwahl entsprechen; das Fehlen einer Handklassifikation erlaubte jedoch lediglich eine Plausibilitätsprüfung. Die beiden folgenden Äußerungen besitzen den größten Abstand zu allen anderen Äußerungen des gesamten Korpus (a) bzw. zu den Äußerungen zur Einleitung eines Überholmanövers (b):

(a) Ach, wie wohl wird's mir um's Herz, wenn ich den arbeitsamen Landmann vor mir sehe. Überhol ihn nicht, herzliebstes Automobil, ich will ihn weitersehen. Ich bin der Chef, kein Mensch wird mich im Betrieb vermissen, ich kann also ruhig zu spät kommen.

(b) Herzliebstes Automobil, wir folgen einfach dem Traktor, haben wir Glück, so führt er uns zum Bäcker meines Vertrauens auf dem Bauernhof. Wenn nicht werde ich im Betrieb anrufen und Frollein Hilde bitten, die Semmeln zu holen.

Sie wurden von offensichtlich unkooperativen Probanden getätigt und konnten durch das Clustering als die am wenigsten typischen Äußerungen identifiziert werden. Im Gegensatz dazu wurde die Äußerung (c) durch

den geringsten Abstand zum Klassenzentrum als typischstes Beispiel für die Einleitung eines Überholmanövers ermittelt:

(c) KoFFI, Überholmanöver starten.

Der zur Erstellung von K2 gewählte Gamification-Ansatz erlaubt keine sinnvolle *textbasierte* Klassifikation, da nur sehr wenige und dazu domänenfremde Wörter gesprochen werden. Abweichend vom ursprünglichen Projektplan wurde daher eine audiobasierte Klassifikation untersucht. Dazu wurden zunächst die vom Spracherkenner verwendeten Folgen spektralen Merkmale (MFCCs) mit rekurrenten neuronalen Netzwerken klassifiziert. Tabelle 2 zeigt die Fehlerraten für die Unterscheidung von „dringenden" und „nicht-dringenden" Äußerungen für verschiedene uni- und bidirektionale Netzwerktypen (LSTM [3, 4] und GRU [5]); die Ergebnisse sind Mittelwerte von 10 unabhängigen Wiederholungen mit zufälliger Initialisierung der Netzwerkparameter.

	LSTM	**GRU**	**BLSTM**	**BGRU**	**DNN**
2 Klassen	33.6	32.3	35.1	33.6	28.9

Tabelle 2: Fehlerraten (in Prozent) für die audio-basierte Unterscheidung von dringenden und nicht-dringenden Äußerungen des Gamification-Korpus.

Das beste Ergebnis wurde hier mit einem einfachen DNN erzielt, wenn jede Äußerung durch einen erweiterten Merkmalvektor charakterisiert wird, der neben den Mittelwerten der MFCCs und Filterbankausgänge auch zusätzliche prosodische Merkmale wie etwa minimale und maximale Grundfrequenz und Energie, Sprechgeschwindigkeit und Pausenlänge enthält. In diesem Fall kam auch eine Oversampling-Technik (*Synthetic Minority Oversampling TEchnique (SMOTE)*, [36]) zum Einsatz, um der hohen a priori Wahrscheinlichkeit der nicht dringenden Äußerungen im Korpus entgegenzuwirken. Eine Unterscheidung der 4 Dringlichkeitsstufen gelang jedoch mit keinem der untersuchten Ansätze.

Gegenüberstellung zur ursprünglichen Planung. Im Bereich der automatischen Spracherkennung ergaben sich keine Abweichungen von der ursprünglichen Planung. Der eingesetzte Spracherkenner konnte um neue Funktionen erweitert werden und die Erkennungsgenauigkeit im automobilen Umfeld konnte durch neue akustische und linguistische Modelle erheblich gesteigert werden. Im Bereich des *Natural Language Understanding* hatte das Fehlen von annotierten Trainingsdaten eine Abweichung von der ursprünglichen Planung zur Folge und die erzielten Ergebnisse bedürfen sicherlich einer zusätzlichen Validierung an öffentlich verfügbaren Korpora. Dennoch konnte Know-How aufgebaut werden und die Arbeiten an den sprach-basierten Interaktionskonzepten innerhalb des Dialogsystems wurden wirkungsvoll unterstützt. Die Bewertung des Zielerreichungsgrades fällt daher insgesamt äußerst positiv aus.

Wichtigste Positionen des zahlenmäßigen Nachweises

Die höchsten Kosten des Teilvorhabens lagen bei den Personalkosten, gefolgt von vorhabenspezifischen Abschreibungen und Reisen. Die über die gesamte Projektlaufzeit angefallenen Kosten in den einzelnen Positionen können dem Verwendungsnachweis entnommen werden.

Notwendigkeit und Angemessenheit der geleisteten Projektarbeit

Die durchgeführten Arbeiten im Teilprojekt waren notwendig und angemessen, da sie dem im Projektantrag formulierten Arbeitsplan entsprachen und alle wesentlichen Aufgaben unter Einhaltung der Ressourcenplanung erfolgreich bearbeitet wurden. Die für Reisen eingeplanten Kosten wurden nicht voll ausgeschöpft. Darüber hinaus waren keine zusätzlichen Ressourcen für das Projekt notwendig.

Voraussichtlicher Nutzen und inhaltliche Verwertbarkeit der Ergebnisse

Die im Bereich der automatischen Spracherkennung untersuchten Modellierungsverfahren werden gegenwärtig auf sämtliche Lösungen und das gesamte Sprachenportfolio von EML übertragen. Dabei steht insbesondere eine weitere Optimierung der zur Laufzeit benötigten Ressourcen im Zentrum, da eine bereits existierende *lokale* on-board Erkennung [37], bei der das Sprachsignal das Fahrzeug nicht verlässt, weiter unterstützt bzw. ausgebaut werden soll.

Eine erste Version der entwickelten online Sprachdetektion wird gegenwärtig in Call-Center-Anwendungen geprüft. Mit Untersuchungen zur Eignung der Algorithmen für die Erkennung von Sprecherwechseln (*„speaker diarization"*) und zur Landessprachenidentifikation (*„spoken language identification"*) wurde im Rahmen der Teilnahme an zwei internationalen Evaluierungen [38, 39] begonnen. Während es für EML generell schwierig ist, seine Produkte bei großen Automobilherstellern oder Zulieferern zu platzieren, zeichnet sich hierfür Interesse bei mittelständischen Unternehmen aus der Call-Center-Branche und im Bereich der öffentlichen Sicherheit ab.

Die Untersuchungen zur (Text-)Klassifikation konnten zum Aufbau von Expertise im Bereich des *Natural Language Understanding* genutzt werden. Das erworbene Know-How soll künftig zur Erweiterung des EML Portfolios im *Speech-Analytics*-Bereich für Call-Center und bei der automatischen Protokollierung von Beratungsgesprächen genutzt werden.

Während der Durchführung des Vorhabens bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Die hier beschriebenen, im Projekt entwickelte und eingesetzte Kernalgorithmen stellen auch am Ende der Projektlaufzeit den Stand der Technik im Bereich der kommerziell verfügbaren Spracherkennungssysteme für den Einsatz im Automobil dar.

Weitere Fortschritte werden derzeit insbesondere von sogenannten *End-to-End-Systemen* erwartet, welche die klassische Trennung und Optimierung von Signalverarbeitung, akustischer Modellierung und linguistischer Modellierung zu Gunsten einer einheitlichen und durchgängigen Modellierung mit neuronalen Netzwerken aufgeben [40]. Es darf angenommen werden, dass derartige Systeme – bei einer radikal einfacheren Systemarchitektur – die Erkennungsleistung bislang eingesetzter Verfahren in den nächsten Jahren übertreffen werden. Ebenso ist ein Zusammenwachsen mit den erwähnten neuronalen Ansätzen zur semantischen Interpretation zu erwarten, die während der Projektlaufzeit entstanden [14, 15].

Vorreiter der skizzierten Entwicklung sind insbesondere große Internet-Konzerne (Amazon, Apple, Facebook, Google), die über genügend Daten und Ressourcen verfügen und aktiv daran arbeiten, ihre jeweiligen Assistenzsysteme (Alexa, Siri, usw.) im Automobil verfügbar zu machen [41].

Über technologische Neuerungen und Konzepte zur sprachlichen Interaktion, die speziell das Szenario des (semi-)autonomen Fahrens adressieren, wurde nichts bekannt.

Erfolgte und geplante Veröffentlichungen des Ergebnisses

Über Weiterentwicklungen des Spracherkennungssystems, die auch im Rahmen von KoFFI durchgeführt wurden, hat EML auf nationalen und internationalen Tagungen berichtet [18, 26, 39] und bei Messeauftritten informiert. Eine gemeinsame Buchpublikation aller in den beiden Bekanntmachungen „Mensch-Technik-Interaktion (MTI) für eine intelligente Mobilität: Verlässliche Technik für den mobilen Menschen" (IMO) und „Individuelle und adaptive Technologien für eine vernetzte Mobilität" (VMO) geförderten Vorhaben wird zur Zeit vorbereitet.

Literatur

1. G. Hinton, L. Deng, D. Yu, et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. In: IEEE Signal Processing Magazine. 29(6), 82–97, 2012.
2. D. Yu and L. Deng (Eds.): Automatic Speech Recognition: A Deep Learning Approach. Springer Verlag, London, 2014.
3. K. Greff, R. L. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber: LSTM: A Search Space Odyssey. In: IEEE Transactions on Neural Networks and Learning Systems. 28(10), 2017.
4. A. Graves, N. Jaitly, and A. Mahamed: Hybrid speech recognition with deep bidirectional LSTM. In: Proc. of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing. Vancouver, Canada, 2013.
5. K. Cho, B. van Merrienboer, Bart, C. Gulcehre, D. Bahdanau, F. Bougares, Fethi, H. Schwenk, and Y. Bengio: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv:1406.1078, 2014.
6. T. Mikolov, K. Chen, G. Corrado, and J. Dean: Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781, 2013.
7. J. Pennington, R. Socher, and C. Manning: GloVe: global vectors for word representation. In: Proc. Of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, 2014.
8. E. Arisoy, T. Sainath, B. Kingsbury, and B. Ramabhadran: Deep Neural Network Language Models. In: Will we ever really replace the N-gram model? On the Future of Language Modelling for HLT. Proc. of the NAACL-HLT 2012 Workshop. Montreal, Canada, 2012.
9. T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur: Recurrent Neural Network based Language Model. In: Proc. of Interspeech 2010. Makuhari, Chiba, Japan, 2010.
10. Y. Goldberg: Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers, San Rafael, CA, USA, 2017.
11. E. Beck, W. Zhou, R. Schlüter, and H. Ney: LSTM Language Models for LVCSR in First-Pass Decoding and Lattice Rescoring. https://arxiv.org/abs/1907.01030, 2019.
12. C. Manning, P. Raghavan, and H. Schütze: Introduction to Information Retrieval. Cambridge University Press, Cambridge, USA, 2008.
13. L. Deng and Y. Liu (Eds.): Deep Learning in Natural Language Processing. Springer Verlag, Singapore, 2018.
14. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin: Attention Is All You Need. https://arxiv.org/abs/1706.03762, 2017.
15. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. Le: XLNet: Generalized Autoregressive Pretraining for Language Understanding. https://arxiv.org/abs/1906.08237, 2019.
16. V. Fischer: Recent Improvements to Neural Network based Acoustic Modeling in the EML Transcription Platform. In: Proc. der Jahrestagung der Deutschen Arbeitsgemeinschaft für Akustik (DAGA). Aachen, Germany, 2016.
17. V. Fischer and S. Kunzmann: Rank based decoding for improved DNN/HMM hybrid Acoustic Models in the EML Transcription Platform. In: Proc. 12. ITG Conference on Speech Communication. Paderborn, Germany, 2016.
18. V. Fischer, O. Ghahabi, and S. Kunzmann: Recent Improvements to Neural Network based Acoustic Modeling in the EML Realtime Transcription Platform. In: Proc. of 29th Conference on Electronic Speech Signal Processing. Ulm, Germany, 2018.
19. P. Doetsch, A. Zeyer, P. Voigtlaender, I. Kulikov, R. Schlüter, H. Ney: *RETURNN: The RWTH Extensible Training framework for Universal Recurrent Neural Networks.* https://arxiv.org/abs/1608.00895, 2016.
20. A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter, and H. Ney: A Comprehensive Study of Deep Bidirectional LSTM RNNs for Acoustic Modeling in Speech Recognition. In: Proc. of the 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing. New Orleans, USA, 2017.
21. A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen: SPEECHDAT-CAR. A Large Speech Database for Automotive Environments. In: Proc. of the 2nd Int. Conf. on Language Resources and Evaluation (LREC). Athens, Greece, 2000.
22. V. Fischer and S. Kunzmann: Bayesian Information Criterion based Multi-style Training and Likelihood Combination for Robust Hands-Free Speech Recognition in the Car. In: Proc. of the IEEE Workshop on Hands-free Speech Communications, Kyoto, Japan, 2001.

23. T. Ko, V. Peddinti, D. Povey, and S. Khudanpur: Audio augmentation for speech recognition. In: Proc. of Interspeech 2015. Dresden, Germany, 2015.
24. S. Watanabe, M. Delcroix, F. Metze, and J. Hershey (Eds.): New Era for Robust Speech Recognition. Exploiting Deep Learning. Springer Verlag, London, 2017.
25. T. He, Y. Fan, Y Qian, T. Tan and K. Yu: Reshaping deep neural network for fast decoding by node-pruning. In: Proc. of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing. Florence, Italy, 2014.
26. O. Ghahabi, W. Zhou, and V. Fischer, A robust voice activity detection for real-time automatic speech recognition. In: Proc. of 29th Conference on Electronic Speech Signal Processing, Ulm, Germany. 2018.
27. European Media Laboratory GmbH: EML Language Model Workplace. Technical Information V 1.7. Heidelberg, 2019.
28. EML European Media Laboratory GmbH: EML Transcription Platform. Architecture, Technical Information, and Services, Version 7.0. Heidelberg, 2015.
29. V. Fischer and S. Kunzmann: The EML Transcription Platform – a flexible transcription environment for robust speech recognition. In: Proc. der Jahrestagung der Deutschen Arbeitsgemeinschaft für Akustik (DAGA), Merano, Italy, 2013.
30. D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, H. Ney: The RWTH Aachen University Open Source Speech Recognition System. In: Proc. of Interspeech 2009. Brighton, UK, 2009.
31. D. Nolden, H. Ney, R. Schlüter: Time Conditioned Search in Automatic Speech Recognition Reconsidered. In: Proc. of Interspeech 2010. Makuhari, Chiba, Japan, 2010.
32. https://www.cs.jhu.edu/~mdredze/datasets/sentiment/index2.html
33. A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts: Learning Word Vectors for Sentiment Analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, 2011.
34. J. Bellegarda: Latent Semantic Mapping. Principles and Applications. Morgan & Claypool Publishers, San Rafael, CA, USA, 2007.
35. A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov: Bag of Tricks for Efficient Text Classification. https://arxiv.org/abs/1607.01759.
36. N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer: SMOTE: Synthetic Minority Over-sampling Technique. In: Journal of Artificial Intelligence Research. Vol. 16, 2002.
37. https://www.pressebox.de/pressemitteilung/european-media-laboratory-gmbh/Sprachsteuerung-fuer-barrierefreies-Fahren-geht-in-Serie/boxid/678529
38. O. Ghahabi and V. Fischer: EML Submission to Albayzin 2018 Speaker Diarization Challenge. In: Proc. IberSPEECH 2018. Barcelona, Spain, 2018.
39. Z. Tang, D. Wang, and Q. Chen: AP18-OLR Challenge: Three Tasks and their Baselines. https://arxiv.org/pdf/1806.00616.pdf
40. S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, T. Ochiai: ESPnet: End-to-End Speech Processing Toolkit. https://arxiv.org/abs/1804.00015
41. S. Kunzmann (Amazon Alexa AI): persönliche Kommunikation, 19.12.2019.

Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart (Schlussbericht oder Veröffentlichung)
geplant	Schlussbericht

3. Titel

Schlussbericht zum Teilvorhaben „Sprachtechnologien für sprachliche Interaktion“ im Verbundprojekt „Kooperative Fahrer-Fahrzeug-Interaktion: Sichere und effiziente Interaktion mit autonomen Fahrzeugen“

4. Autor(en) [Name(n), Vorname(n)]	5. Abschlussdatum des Vorhabens
Volker Fischer	31.10.2019
	6. Veröffentlichungsdatum 30.04.2020
	7. Form der Publikation Schlussbericht

8. Durchführende Institution(en) (Name, Adresse)	9. Ber. Nr. Durchführende Institution
EML European Media Laboratory GmbH Berliner Strasse 45 D – 69120 Heidelberg	10. Förderkennzeichen 16SV7627
	11. Seitenzahl 11

12. Fördernde Institution (Name, Adresse)	13. Literaturangaben
Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	41
	14. Tabellen 2
	15. Abbildungen 0

16. Zusätzliche Angaben

17. Vorgelegt bei (Titel, Ort, Datum)

Die Arbeiten wurden teilweise publiziert bei:
ESSV2018: 29. Konferenz für Elektronische Sprachverarbeitung, Ulm, März 2018.

18. Kurzfassung

Im Bereich von automatischer Spracherkennung und NLP haben (rekurrente) neuronale Netzwerke klassische Machine Learning Techniken wie etwa Hidden Markov Modelle oder Entscheidungsbäume weitgehend verdrängt. Bidirektionale Long Short Term Memory Netzwerke (BLSTMs) wurden hier zur akustisch-linguistischen Modellierung eingesetzt. Dabei entstand ein Spracherkennungssystem, das auch eine online-fähige voice activity detection umfasst. Es ist Teil eines innovativen Sprachdialogsystems für (teil-)autonome Fahrzeuge und ermöglicht es dem Dialogsystem, mehr als 95 Prozent aller Nutzeräußerungen korrekt zu interpretieren. Dabei kommt auch ein Klassifikator zum Einsatz, der dringende von weniger dringenden Anfragen unterscheidet.
Das entwickelte System unterstützt damit die Entwicklung eines ganzheitlichen, psychologisch fundierten Konzepts zur kooperativen Fahrer-Fahrzeug-Interaktion für hochautomatisierte Fahrzeuge.

19. Schlagwörter

Spracherkennung, Neuronale Netze, NLP, autonomes Fahren

20. Verlag	21. Preis
--	--