EVALUIERUNG DER LEISTUNGSFÄHIGKEIT VON SPRACHERKENNUNGSSYSTEMEN IN GERÄTEN DER KOMMUNIKATIONSTECHNIK

Andreas Kitzig und Hans-Günter Hirsch

iPattern – Institut für Mustererkennung, Hochschule Niederrhein, 47805 Krefeld
{andreas.kitzig, hans-guenter.hirsch}@hs-niederrhein.de

Kurzfassung: Bei der Entwicklung von Spracherkennungssystemen, die in einer gestörten akustischen Umgebung eingesetzt werden sollen, wird zur Evaluation der Spracherkennungsleistung eine Vielzahl an Erkennungsexperimenten durchgeführt. Dabei werden im Rahmen von Simulationsexperimenten ungestörte und gestörte Sprachaufnahmen aus unterschiedlichen Störumgebungen verwendet. Häufig wird die Erkennungsleistung der entwickelten Systeme mit der Erkennungsleistung bereits vorhandener Systeme verglichen. Diese Schritte zur Evaluation werden durchgeführt, um einen Überblick über die Leistungsfähigkeit der entwickelten Systeme in den einzelnen Störszenarien zu erlangen und eine Einordnung im Vergleich zu den bereits vorhandenen Systemen vornehmen zu können. In der vorgestellten Untersuchung wird die Leistung zweier im Rahmen früherer Arbeiten entwickelter, robuster Erkennungssysteme mit der Leistungsfähigkeit von Erkennungssystemen verglichen, die als integraler Bestandteil einiger kommerziell verfügbarer Geräte aus dem Bereich der Kommunikationstechnik eingesetzt werden. Eines der beiden im Labor entwickelten Erkennungssysteme basiert auf einer Extraktion robuster akustischer Merkmale, das andere basiert auf einer Adaption der Referenzmuster auf die jeweiligen akustischen Bedingungen bei der Spracheingabe. Bei den Geräten aus dem Bereich der Kommunikationstechnik handelt es sich um Systeme, die über eine Option zur Spracherkennung von Ziffern und Ziffernketten zum Aufbau einer Telefonverbindung verfügen. Zur Durchführung von Experimenten mit den kommerziell verfügbaren Geräten wurde ein Testaufbau entwickelt, mit dem man in einem reflexionsarmen Raum die Worterkennungsraten bei gesprochenen Folgen von Ziffern bestimmen kann. Zuerst wird eine Übersicht über die selbstentwickelten robusten Erkennungssysteme gegeben. Anschließend werden die verwendeten Geräte und deren Funktionsumfang sowie der Testaufbau der Experimente und die Testdaten vorgestellt. Abschließend erfolgt die Darstellung der Ergebnisse der Tests sowie deren Bewertung.

1 Robuste Verfahren zur Spracherkennung

Es werden die beiden Verfahren zur robusten Spracherkennung vorgestellt, die im Rahmen früherer Arbeiten entwickelt wurden [1],[2]. Die meisten Erkennungssysteme zeigen einen deutlichen Einbruch der Erkennungsleistung, wenn die akustischen Bedingungen bei der Aufnahme der Sprachdaten, die zum Training verwendet werden, nicht mit den Bedingungen beim Einsatz des Systems übereinstimmen. Daher hat man verschiedene Ansätze entwickelt, um diese Verschlechterung der Erkennung bei unterschiedlichen akustischen Bedingungen zu reduzieren.

Ein einfacher Ansatz ist die Aufnahme oder Erzeugung von Sprachsignalen, die alle zu erwartenden akustischen Bedingungen beinhalten. Dies setzt allerdings die Kenntnis der akustischen Umgebungen, in denen das System eingesetzt werden soll, voraus. Mit Hilfe dieser Signale können Referenzmodelle erzeugt werden, die man als so genannte "Multi-Condition" Modelle bezeichnet. Damit kann man über alle Bedingungen hinweg eine recht gute mittlere Erkennung gewährleisten.
Ansonsten kann man eine grobe Einteilung der meisten robusten Erkennungsverfahren in zwei Klassen vornehmen. Der erste Ansatz beruht auf einer Sprachanalyse, bei der robuste akustische Merkmale gewonnen werden, die möglichst unabhängig von der akustischen Aufnahmesituation sind. Die zweite Klasse von Verfahren verwendet eine Anpassung der akustischen Merkmale, die in den Referenzmustern enthalten sind, auf die jeweilige Störsituation. Die beiden im Folgenden näher vorgestellten Verfahren gehören jeweils einer der beiden Kategorien an. In beiden Fällen werden als akustische Merkmale 12 Cepstralparameter und ein Energiekoeffizient sowie deren erste und zweite Ableitungen, die so genannten Delta und Delta-Delta Koeffizienten, verwendet. Insgesamt sind es 39 Merkmalswerte, die aus einer Analyse kurzer Sprachsegmente mit einer Dauer von etwa 25 ms und in einem zeitlichen Abstand von 10 ms hervorgehen.
Bei dem Verfahren zur Extraktion robuster Merkmale [2] wurde die Cepstralanalyse um eine adaptive Filterung im Spektralbereich ergänzt, mit der Hintergrundstörgeräusche reduziert werden sollen. Die adaptive Filterung beruht auf einer Schätzung des Spektrums der Hintergrundstörung und verwendet als speziellen Verarbeitungsschritt eine Glättung von Cepstralparametern in zeitlicher Richtung [5],[6] zur Bestimmung der Filtercharakteristik. Dieses Verfahren wird im Weiteren mit Hilfe des Kürzels "HGH-robust" referenziert. Bei dem Adaptionsverfahren [1] werden die Mittelwerte, die die Gauß-Verteilungen der akustischen Merkmale in den Zuständen eines statistisch basierten Hidden-Markov Modells festlegen, an die jeweiligen akustischen Bedingungen bei einer Spracheingabe angepasst. Dazu wird ebenfalls das Spektrum der Hintergrundstörung geschätzt. Zudem erfolgt eine Schätzung der Nachhallzeit, mit der grob der Einfluss eines Raumes im Fall einer Spracheingabe im Freisprechmodus beschrieben werden kann. Das Adaptionsverfahren nimmt eine Anpassung an die Störsituation und an den Hall des Raumes vor. Das Kürzel "HGH-adapt" wird zur Referenzierung dieses Verfahrens verwendet.

2 Testgeräte

Es wurden insgesamt vier verschiedene Geräte getestet. Hierbei handelt es sich um aktuelle, kommerziell erhältliche Systeme von unterschiedlichen Herstellern. Es wurden Geräte ausgewählt, die eine Erkennung von Ziffernketten und Einzelziffern zur Eingabe von Telefonnummern per Spracherkennung ermöglichen. Die Auswahl fiel auf zwei Freisprechanlagen zum Festeinbau im KFZ und zwei mobile Geräte. An dieser Stelle ist anzumerken, dass die Auswahl an geeigneten Geräten relativ gering ist, da nur wenig eine erweiterte Spracherkennung mit Ziffernwahl bieten. Die meisten derzeit verfügbaren Geräte beschränken sich auf die Spracherkennung von Telefonbucheinträgen. Einige der verwendeten Geräte verfügen zwar über die zusätzliche Möglichkeit zur Steuerung von Funktionen per Sprachkommando, jedoch wurde in dieser Untersuchung nur die Ziffernerkennung getestet. Nachfolgend werden die für die Tests verwendeten Geräte kurz vorgestellt:

- Bury CC9060 Plus
- Funkwerk Dabendorf (FWD) Ego Flash
- Falk Navigationsgerät F10

- Apple iPhone 3Gs

Das *Bury CC9060 Plus* ist als Gerät für den Festeinbau im KFZ konzipiert. Es ermöglicht eine Spracherkennung von Kommandowörtern und Ziffern ohne vorheriges Training. Als Option verfügt das Gerät über die Möglichkeit zur Sprachwahl von Telefonnummern sowie zur Aktivierung der Sprachsteuerung durch ein so genanntes "Zauberwort". Weiterhin kann per Sprachbefehl eine Rufannahme/Ablehnung durchgeführt werden.
Bei dem zweiten Gerät *Funkwerk Dabendorf (FWD) Ego Flash* handelt es sich ebenfalls um ein Gerät, das zum Festeinbau im KFZ konzipiert ist. Der Funktionsumfang ist dem Bury-Gerät sehr ähnlich, auch hier besteht die Möglichkeit zur Sprachsteuerung und Sprachwahl ohne vorheriges Training, wobei auch hier Kommandowörter und Ziffern erkannt werden können. Das dritte Gerät ist ein mobiles Navigationssystem von *Falk* mit der Bezeichnung *F10*. Bei diesem Gerät kann die Navigationsanwendung per Sprache gesteuert werden. Zusätzlich wurde eine Freisprechfunktion inkl. Sprachwahl integriert, die eine Erkennung von Kommandowörtern und Ziffern bietet. Das Gerät ist ohne vorheriges Training nutzbar. Das vierte Gerät ist ein *Apple iPhone 3Gs*. Dabei handelt es sich um ein Smartphone, das neben der Sprachsteuerung des Mediaplayers mit Hilfe von Kommandowörter auch über eine Freisprechfunktion mitsamt Sprachwahl zur Erkennung von Kommandowörtern und Ziffern verfügt. Auch hier ist kein vorheriges Training notwendig. Das iPhone ist im Gegensatz zu den drei anderen Geräten nicht ausschließlich für die Verwendung im KFZ konzipiert und optimiert worden. Daher sind die in den Tests erzielten Ergebnisse nicht direkt mit den Ergebnissen der anderen Geräte vergleichbar.

3 Testaufbau

In diesem Abschnitt wird der Testaufbau erläutert. Die Gerätetests wurden alle im reflexionsarmen Raum des Akustiklabors der Hochschule Niederrhein durchgeführt. Dadurch konnte sicher gestellt werden, dass die verwendeten gestörten halligen Sprachdaten nicht durch zusätzliche Einflüsse des Raums, in dem die Wiedergabe erfolgt, verfälscht werden. Um die Durchführung der Experimente zu vereinfachen und Bedienfehlern vorzubeugen, wurde ein halbautomatischer Versuchsaufbau erstellt. Hierbei bedeutet halbautomatisch, dass die Bedienperson den Test für jede Äußerung neu initialisiert und abschließend das Erkennungsergebnis notiert, da diese Schritte nicht ohne erheblichen Mehraufwand automatisiert werden konnten. Der Aufbau ist derart gestaltet, dass die Tests von einer einzelnen Person durchgeführt werden können. Eine Ausnahme stellt das iPhone 3Gs dar. Hier wird eine weitere Person im reflexionsarmen Raum zur Initialisierung der einzelnen Testschritte benötigt. Der Testaufbau ist als Blockschaltbild in Abbildung 1 dargestellt. Durch den Testaufbau in Abbildung 1 wird die Interaktion eines

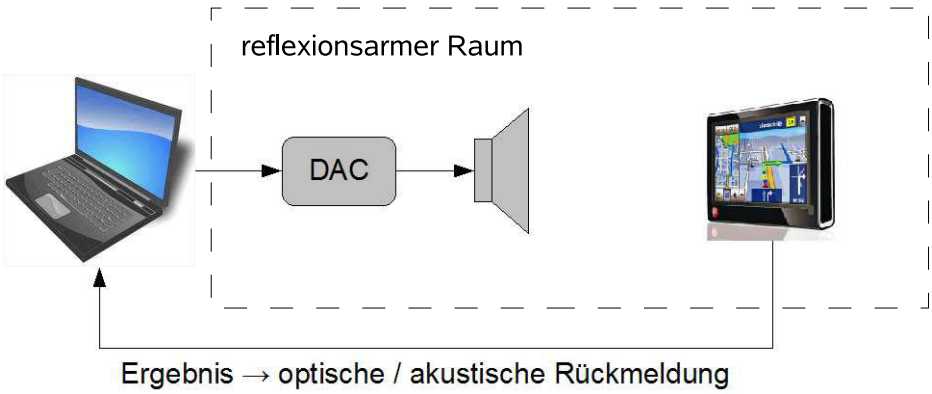

Abbildung 1 - Blockschaltbild des Testaufbaus

Benutzers mit dem Gerät in einer gestörten Umgebung, z.B. im „alltäglichen“ Einsatz im KFZ, simuliert. Neben dem Testgerät, für das in der Abbildung beispielhaft das Falk Navigationsgerät steht, kommt ein Notebook mit externer USB-Soundkarte und ein Aktivlautsprecher zur Wiedergabe der Testdaten zum Einsatz. Das Testgerät und der Aktivlautsprecher befinden sich im reflexionsarmen Raum, die Dialogsteuerung per Notebook und die Bedienung des entsprechenden Testgerätes durch ein abgesetztes Bedienteil erfolgt von außerhalb. Allgemein setzt sich ein möglicher Sprachdialog aus mehreren Schritten zusammen, wobei einzelne Details je nach Gerät abweichen können. Der Dialog zwischen Benutzer und Gerät sowie die daraus abgeleitete Vorgehensweise zur Simulation sind in Tabelle 1 dargestellt. Zur Durchführung der Tests

Tabelle 1 - Übersicht Dialoge

realer Dialog	**nachgebildeter Dialog für die Experimente**
- Gerät mittels Tastendruck in den Erkennungsmodus versetzen	
- Kommandowort zur Initialisierung der Ziffernwahl sprechen (z.B. “wählen” etc.) und auf Quittierungston warten	- initiales Sprachkommando zum Wählen der Nummer abspielen und auf Quittierungston warten
- Ziffernfolge sprechen	- Testäußerung abspielen
- Erkennung der Ziffernfolge durch das Gerät	
- Erkennungsergebnis prüfen (Blick auf das Display und/oder anhören der akustischen Rückmeldung)	- Erkennungsergebnis erfassen und notieren
- Ergebnis im Gerät mittels Sprachkommando bestätigen oder ggf. löschen	- Sprachkommando zum Löschen des Ergebnis im Gerät abspielen

steht der Bedienperson eine grafische Benutzeroberfläche (siehe Abbildung 2) zur Verfügung, über die die Liste mit den Testdaten und die Kommandowörter zur Initialisierung einzeln verarbeitet werden können. Nach erfolgter Verarbeitung und Erkennung der jeweiligen Testäußerung

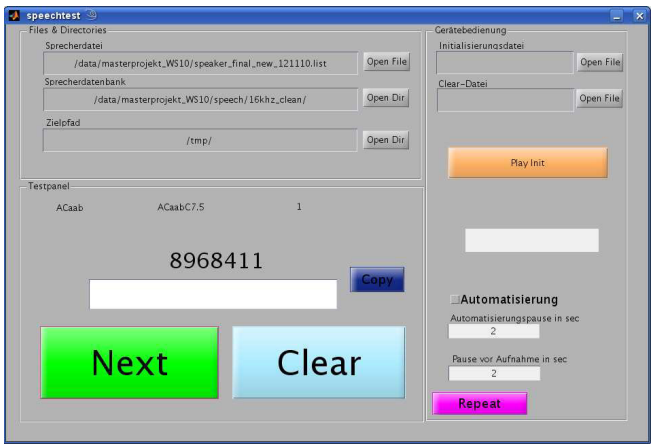

Abbildung 2 - Benutzeroberfläche

kann die Rückmeldung des Testgerätes, die das Erkennungsergebnis enthält, von der Bedienperson in der Benutzeroberfläche dokumentiert werden. Anschließend wird die nächste Äußerung abgespielt, bis die vollständige Liste verarbeitet wurde. Die Rückmeldung erfolgt bei allen Testgeräten sowohl optisch als Anzeige auf dem Bildschirm des Testgerätes als auch akustisch mittels Sprachsynthese und ist somit für die Bedienperson ohne großen Aufwand zugänglich.

4 Testdaten

Nach der Erläuterung zum Testaufbau werden in diesem Abschnitt die Spracheingaben zum Test der Verfahren vorgestellt. Die Testdaten werden für alle Tests, d.h. für die Tests mit den vier kommerziellen Geräten und für die Tests mit den eigenen Systemen, verwendet. Um eine möglichst realitätsnahe Testumgebung zu schaffen, wurde die Eingabe einer Ziffernkette im Freisprechmodus in einem fahrenden KFZ simuliert, da alle getesteten Geräte für eine Anwendung in dieser Störumgebung konzipiert und vermutlich dafür optimiert wurden. Dabei wurde eine Kette aus sieben Ziffern, die eine Telefonnummer darstellen soll, als Testäußerung akustisch wiedergegeben. Die Begrenzung auf sieben Ziffern war nötig, um gleiche Testbedingungen zu gewährleisten, da einige Testgeräte nur Blöcke von maximal sieben Ziffern verarbeiten können. Als Basis wurden ungestörte Daten einer eigenen Sprachdatenbank in deutscher Sprache verwendet, wobei das Grundexperiment auf folgenden Daten basiert:

- 100 Sprachäußerungen
- 50 verschiedene Sprecher (25 männliche und 25 weibliche Sprecher)
- 7 Ziffern pro Äußerung
- Ziffern von 0 bis 9 ohne zwo

Unter Verwendung der Grunddaten sowie unterschiedlicher "KFZ- bzw. Alltags- Störgeräuschen" und Raumimpulsantworten wurden mit Hilfe eines eigenen Tools "SIREAC" [3] gestörte Daten für die Tests erzeugt. SIREAC ermöglicht die Überlagerung von Sprachsignalen mit Störgeräuschen in unterschiedlichen SNRs, die Faltung mit Raumimpulsantworten, sowie die Möglichkeit zur Simulation von verschiedenen Übertragungskanälen. Das Tool steht als Web-Demo [3] für eigene Experimente zur Verfügung. Für die Tests wurden insgesamt sechs Versionen der Testdaten generiert:

1. CarNoise xxdB HF (KFZ Hintergrundstörung (SNR 5, 10, 15, und 20dB) und Raumimpulsantwort 1)
2. CarNoise 10dB HF BS (KFZ Hintergrundstörung (SNR 10dB) und Raumimpulsantwort 2)
3. IntNoise 10dB HF (Störgeräusche aus Alltagssituationen (SNR 10dB) und Raumimpulsantwort 3)

Die gestörten "CarNoise"-Daten (Datensätze 1 und 2) repräsentieren eine Spracheingabe im Freisprechmodus in einem fahrenden KFZ in unterschiedlichen Signal-Rausch-Verhältnissen (SNR) und Sitzpositionen (Raumimpulsantwort 1 und 2). Hierbei wird durch die Verwendung von Raumimpulsantwort 1 eine Spracheingabe im Freisprechmodus vom Kopf des Fahrers zum Spiegel simuliert, durch die Verwendung von Raumimpulsantwort 2 wird eine Spracheingabe im Freisprechmodus vom Kopf einer Person auf dem hinteren linken Rücksitz zum Spiegel simuliert.
Die Endposition des Spiegels im Bezug zum Sprecher in den verwendeten Raumimpulsantworten wurde gewählt, da dieser Bereich im KFZ oft universell genutzt wird. Zum einen wird das Mikrofon von Freisprechsystemen zum Festeinbau häufig am Spiegel montiert und zum anderen werden mobile Geräte meist in einer ähnlichen Position in unmittelbarer Nähe zum Spiegel an der Frontscheibe befestigt. Somit wird sichergestellt, dass die Testdaten für alle Testsysteme

verwendet werden können. Die gestörten “InteriorNoise” Daten, kurz “IntNoise”, entsprechen einer Spracheingabe im Freisprechmodus in einer gestörten Büroumgebung. Die verwendete Raumimpulsantwort 3 beschreibt hierbei den Abstand vom Kopf einer Person, die an einem Schreibtisch in einem mittelgroßen Büro (Nachhallzeit $T_{60} \approx 0,7s$) sitzt, zu einem Mikrofon in ca. 70cm Abstand, das auf dem Schreibtisch steht.
Im Gegensatz zu den als stationär zu betrachtenden CarNoise-Hintergrundstörungen ist die IntNoise-Störumgebungen geprägt von nicht stationären Störgeräuschen und zusätzlichen Sprechern im Hintergrund. Der Datensatz 3 wurde für einen zusätzlichen Test erzeugt, um die Reaktion der Systeme auf unbekannte Störumgebungen zu untersuchen. Zu diesem Zweck eignet sich besonders die mit InteriorNoise gestörten Daten, da sie eine erhebliche Abweichung von den KFZ-Störumgebungen darstellen und eine Erkennung durch den speziellen Aufbau der Hintergrundstörung erschweren.
Neben den gestörten Daten wurden die ungestörten Ausgangsdaten (Testumgebung „Clean“) als Referenz verwendet.

5 Ergebnisse

Nachfolgend werden die Ergebnisse der Experimente in den Tabellen 2 und 3 in Form von Wort-Erkennungsraten dargestellt. Die Auswertung der Ergebnisse erfolgte mit dem Tool “HResults” des HTK Speech Recognition Toolkit [4].
Zu den Experimenten ist anzumerken, dass alle vier kommerziellen Geräte mit einer Funktion zur Rückweisung arbeiten. Das bedeutet, dass die Geräte eine Äußerung, die nicht sicher verarbeitet werden kann, durch eine Sprachausgabe zurückweisen, um eine fehlerhafte Erkennung zu vermeiden. Eine mögliche Sprachausgabe zur Rückweisung ist z.B.: “Ich habe sie nicht verstanden.”. Die Rückweisung hat zur Folge, dass sich die Anzahl an erkannten Äußerungen bei den verschiedenen Experimenten unterscheidet. Die Rückweisungsrate liegt je nach verwendetem Gerät bei den ungestörten Daten und den Daten mit KFZ Hintergrundstörung zwischen 0% und 15%, das iPhone 3Gs weist hierbei Rückweisungsraten bis zu 29% auf. Bei der InteriorNoise-Störumgebung kommt es je nach verwendetem Gerät zu Rückweisungsraten von 12% bis zu 85% .

Tabelle 2 - Übersicht Ergebnisse Testgeräte

	Bury CC9060+	*FWD Ego Flash*	*Falk F10*	*Apple iPhone 3Gs*
Clean	97,86 %	93,43 %	95,45 %	89,07%
CarNoise 20dB HF	96,14 %	95,96 %	95,08 %	88,30%
CarNoise 15dB HF	94,61 %	95,14 %	93,98 %	74,29%
CarNoise 10dB HF	92,03 %	94,20 %	93,92 %	46,14%
CarNoise 10dB HF BS	88,44 %	88,89 %	88,39 %	42,15%
CarNoise 5dB HF	81,34 %	78,90 %	81,27 %	27,16%
IntNoise 10dB HF	65,48 %	68,18 %	50,48 %	7,68%

In Tabelle 2 sind die Ergebnisse der verschiedenen Experimente, die mit den kommerziellen Testgeräten durchgeführt wurden, dargestellt. Werden die Erkennungsraten der Systeme miteinander verglichen, wird deutlich, dass die Geräte Bury CC9060 Plus, FWD Ego Flash und Falk F10 eine ähnliche Erkennungsleistung in den einzelnen KFZ Störumgebungen zeigen, die mit abnehmendem SNR schlechter wird. Weiterhin geht aus den Ergebnissen in Tabelle 2 hervor, dass keines der Geräte bei Tests mit ungestörten Clean- Daten eine für diese Störumgebung

typische Worterkennungsrate von annähernd 99% erreicht. Zusätzlich ist ein erheblicher Einbruch der Erkennungsrate bei Verwendung der InteriorNoise Testdaten zu verzeichnen.
Die Auswertung der Ergebnisse bestätigt die eingangs aufgestellte Vermutung, dass die drei Geräte für eine Verwendung in einer KFZ Störumgebung optimiert wurden, um eine möglichst hohe Erkennungsleistung bzw. eine möglichst fehlerfreie Erkennung zu garantieren. Andere Umgebungen, z.B. ungestörte Umgebungen oder gestörte Umgebungen mit Alltagsgeräuschen, die bei der Anwendung der Geräte keine Bedeutung haben, werden vernachlässigt. Da über die technische Realisierung der Erkennungsysteme nichts bekannt ist, kann aus der Betrachtung der Ergebnisse gefolgert werden, dass in den Geräten Multi- Condition- Modelle zur Spracherkennung verwendet werden, die speziell für die KFZ Störsituation trainiert wurden. Der Einsatz eines robusten Verfahrens in den Testgeräten ist eher unwahrscheinlich, da sich sonst bei den InteriorNoise Testdaten und bei den ungestörten „Clean“- Daten bessere Ergebnisse hätten einstellen müssen.
Werden abschließend die Ergebnisse der drei Geräte mit denen des iPhone verglichen, kann festgestellt werden, dass das iPhone sowohl bei den Tests in gestörten Umgebungen als auch bei dem Test mit ungestörten Daten wesentlich schlechtere Erkennungsraten liefert. Die Spracherkennungsfunktion in dem Gerät wurde vermutlich nicht für den Einsatz in einer speziellen Umgebung optimiert.
Um nun den Vergleich der Ergebnisse der kommerziellen Systeme mit den Ergebnissen der selbstentwickelten robusten Verfahren zu ermöglichen, werden in Tabelle 3 die Ergebnisse für das FWD Ego Flash zusammen mit den Ergebnissen der beiden selbstentwickelten robusten Verfahren dargestellt. Hierbei wurde bei der Auswertung der Erkennungsergebnisse die von dem Gerät Ego Flash zurückgewiesenen Äußerungen nicht berücksichtigt, um eine Vergleichbarkeit der Erkennungsraten zu gewährleisten. Die Ergebnisse in Tabelle 3 sind exemplarisch für alle Testsysteme dargestellt, da sich die Rückweisungsraten der kommerziellen Systeme und somit die Ergebnisse der Experimente mit den HGH-Systemen nur geringfügig für die einzelnen Systeme unterscheiden.
Nachdem sich in den Tests herausgestellt hat, dass in den kommerziellen Geräten vermutlich Multi- Condition- Modelle eingesetzt werden, wurden weitere Experimente mit dem Verfahren “HGH robust” unter Verwendung von auf die KFZ Störsituation trainierten Multi- Condition Modellen durchgeführt.
Werden nun die Ergebnisse der eigenen Systeme mit denen des kommerziellen Systems ver-

Tabelle 3 - Übersicht Ergebnisse FWD Ego Flash

	FWD Ego Flash	*HGH adapt*	*HGH robust*	*HGH rob multi*
Clean	93,43%	99,43%	99,86%	99,00%
CarNoise 20dB HF	95,96%	96,97%	97,69%	98,41%
CarNoise 15dB HF	95,14%	95,00%	94,00%	97,71%
CarNoise 10dB HF	94,20%	90,18%	87,95%	95,98%
CarNoise 10dB HF BS	88,89%	87,16%	84,99%	92,06%
CarNoise 5dB HF	78,90%	77,27%	73,21%	89,77%
IntNoise 10dB HF	68,18%	70,13%	66,07%	58,44%

glichen, wird deutlich, dass die beiden robusten Systeme “HGH adapt” und “HGH robust” in der ungestörten Testumgebung (“Clean”) und bei gestörten Daten im oberen SNR Bereich bessere Ergebnisse liefern. Im unteren SNR Bereich sind die Erkennungsraten etwas schlechter im Vergleich zu den Testgeräten. Die geringfügig besseren Ergebnisse im unteren SNR-Bereich bei den kommerziellen Geräten liegen in der Verwendung der auf die Störsituation angepass-

ten Modelle begründet. Diese nahezu optimale Anpassung kann mittels der robusten Verfahren nicht erreicht werden. Dafür sind die selbstentwickelten Systeme universeller anwendbar, was z.B. die bessere Erkennungsrate bei den “Clean”- Testdaten zeigt. Eine Erweiterung des Systems HGH robust um Multi- Condition-Modelle (HGH rob. multi) führt in nahezu allen Fällen zu einer Steigerung der Erkennungsrate im Vergleich zu den Testsystemen. Jedoch stellt die Erzeugung von auf die Störsituation trainierten Modellen einen erheblichen Mehraufwand dar, so dass hierbei Aufwand und Nutzen im Vergleich zu den Ergebnissen der robusten Basissysteme geprüft werden muss.

6 Zusammenfassung

Im Rahmen der hier vorgestellten Untersuchungen wurde ein Vergleich der Leistungsfähigkeit von kommerziell verfügbaren Geräten aus dem Bereich der Kommunikationstechnik, die eine Bedienung mit Hilfe einer Spracherkennung erlauben und zwei im Labor entwickelten, robusten Spracherkennungssystemen angestellt. Dazu wurden vier kommerzielle Geräte unter Verwendung von ungestörten und gestörten Sprachdaten mittels eines zuvor festgelegten halbautomatischen Versuchsaufbaus im reflexionsarmen Raum getestet. Zusammenfassend ergibt sich aus der Betrachtung der Ergebnisse, dass die selbstentwickelten Systeme im Vergleich zu den kommerziellen Systemen eine vergleichbar gute Leistungsfähigkeit aufweisen, ohne auf die spezielle Störsituation angepasst zu werden. Erfolgt noch eine zusätzliche Anpassung der Modelle auf die Störumgebung, lässt sich noch eine deutliche Verbesserung erzielen.

Literatur

[1] H.G. HIRSCH, H. FINSTER, A New Approach for the Adaptation of HMMs to Reverberation and Background Noise *Speech Communication, Vol.50, S. 244-263, 2008*

[2] H. G. HIRSCH, A. KITZIG, Robust Speech Recognition by Combining a Robust Feature Extraction with an Adaptation of HMMs *9. ITG Fachtagung Sprachkommunikation, 2010*

[3] H.G. HIRSCH, H. FINSTER, The Simulation of Realistic Acoustic Input Scenarios for Speech Recognition Systems *EuroSpeech, 2005*, Web-Demo unter *http://dnt.kr.hsnr.de*

[4] HTK SPEECH RECOGNITION TOOLKIT, Cambridge University Engineering Department, *htk.eng.cam.ac.uk*

[5] C. BREITHAUPT, R. MARTIN, DFT based speech enhancement for robust automatic speech recognition *ITG Fachtagung Sprachkommunikation, Aachen, 2008*

[6] C. BREITHAUPT, T. GERKMANN, R. MARTIN, Cepstral smoothing of spectral filter gains for speech enhancement without musical noise *IEEE Signal processing letters, 2007*