Abschlussbericht

des Projekts mit dem Thema

Verbesserung der automatischen Erkennung gestörter Sprachsignale durch Verwendung robuster akustischer Merkmale und einer Adaption der Referenzmuster

Berichtszeitraum: April 2004 bis Juni 2008

Juli 2008

Prof. Dr.-Ing. Hans-Günter Hirsch

Hochschule Niederrhein

Fachbereich Elektrotechnik und Informatik

Hochschule Niederrhein

Niederrhein University of Applied Sciences

1. Allgemeine Angaben

1.1. DFG-Geschäftszeichen

HI 384/1-4

1.2. Antragsteller

Prof. Dr. Ing. Hans-Günter Hirsch
Fachbereich Elektrotechnik und Informatik
Hochschule Niederrhein

1.3. Thema

Verbesserung der automatischen Erkennung gestörter Sprachsignale durch Verwendung robuster akustischer Merkmale und einer Adaption der Referenzmuster

1.4. Berichtszeitraum

Das gesamte Projekt wurde im Zeitraum vom 01.04.2004 bis zum 30.06.2008 von der DFG gefördert. Die Durchführung und Förderung erfolgte dabei in 2 Teilprojekten. Das erste Teilprojekt wurde in der Zeit vom 01.04.2004 bis zum 30.09.2005 bearbeitet. Die Ergebnisse wurden bereits im Fortsetzungsantrag dokumentiert. Es erfolgt nochmals eine Zusammenfassung der wesentlichen Erkenntnisse im nachfolgenden Abschnitt. Das zweite Teilprojekt wurde in der Zeit vom 01.10.2005 bis zum 30.06.2008 bearbeitet. Die Ergebnisse dieses Teilprojekts werden in etwas ausführlicherer Form wiedergegeben.

1.5. Liste der Publikationen

[1] H.G. Hirsch: Automatic Speech Recognition in Adverse Acoustic Conditions, Kapitel des Buchs *Advances in Digital Speech Transmission*, Herausgeber: R. Martin, U. Heute, C. Antweiler, Verlag John Wiley & Sons, S. 461-496, Januar 2008

[2] 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, März 2008

[3] H.G. Hirsch: Automatic Speech Recognition in Adverse Acoustic Conditions, Buch in der Schriftenreihe des Fachbereichs Elektrotechnik und Informatik, Hochschule Niederrhein, Shaker Verlag, Mai 2008

[4] H.G. Hirsch, H. Finster: “The Simulation of Realistic Acoustic Input Scenarios for Speech Recognition Systems”, *Interspeech conference 2005*, S. 2697-2700, Lissabon, Portugal, 2005.

[5] H.G. Hirsch, H. Finster; A New HMM Adaptation Approach for the Case of a Hands-free Speech Input in Reverberant Rooms, *Interspeech conference 2006*, S. 2697-2700, Pittsburgh, USA

[6] H.G. Hirsch, P. Pogscheba: Verbesserung der Spracherkennung bei Freisprechen durch eine robuste Merkmalsextraktion und eine Adaption der Referenzmuster, wird auf der im September 2008 stattfindenden ITG Fachtagung Sprachkommunikation veröffentlicht

[7] H.G. Hirsch, A. Kitzig: Visualisierung der in einem HMM enthaltenen spektralen Merkmale, wird auf der im September 2008 stattfindenden ITG Fachtagung Sprachkommunikation veröffentlicht

[8] H. Wang, D.Gelbart, H.G. Hirsch, W. Hemmert: The Value of Auditory Offset Compensation and Appropriate Acoustic Modeling, wird auf der *Interspeech conference* im September 2008 in Brisbane, Australien veröffentlicht werden

[9] H. Finster, H.G. Hirsch: Sprachdatenbasis „Aurora-5“, Sammlung künstlich gestörter und in verhallter Umgebung aufgenommener Sprachdaten, die im Rahmen des Projekts erzeugt wurde, erhältlich bei ELRA (European Language Ressource Association), 2007

2. Arbeits- und Ergebnisbericht

2.1. Zielsetzung

Die Zielsetzung des Vorhabens war die Untersuchung und Entwicklung von Verfahren zur Verbesserung einer automatischen Erkennung von Sprache bei einer Spracheingabe in einer gestörten und verhallten Umgebung. Dazu wurde ein neuer Ansatz zur Adaption der bei der Spracherkennung verwendeten Referenzmuster auf den Einfluss des Nachhalls entwickelt, der mit einem bereits vorhandenen Verfahren zur Adaption auf Hintergrundstörungen kombiniert werden kann. Zudem wurden verschiedene Möglichkeiten der Kombination einer Extraktion robuster akustischer Merkmale mit einer Adaption der Referenzmuster untersucht.

2.2. Durchgeführte Arbeiten

Die zur Erreichung des gesetzten Ziels durchgeführten Arbeiten fanden in vier Abschnitten statt.

In der ersten Projektphase wurden die akustischen Bedingungen bei der Spracheingabe und Sprachübertragung, die beim praktischen Einsatz eines Spracherkennungssystems auftreten, analysiert und mit Hilfe eines Softwarewerkzeugs simuliert. Die Simulation beinhaltet, wie es auch Bild 1 veranschaulicht,

- die Aufnahme in einer verhallten Umgebung im Freisprechmodus,
- die additive Überlagerung einer Hintergrundstörung sowie
- die mögliche Übertragung des Sprachsignals über einen Mobilfunkkanal.

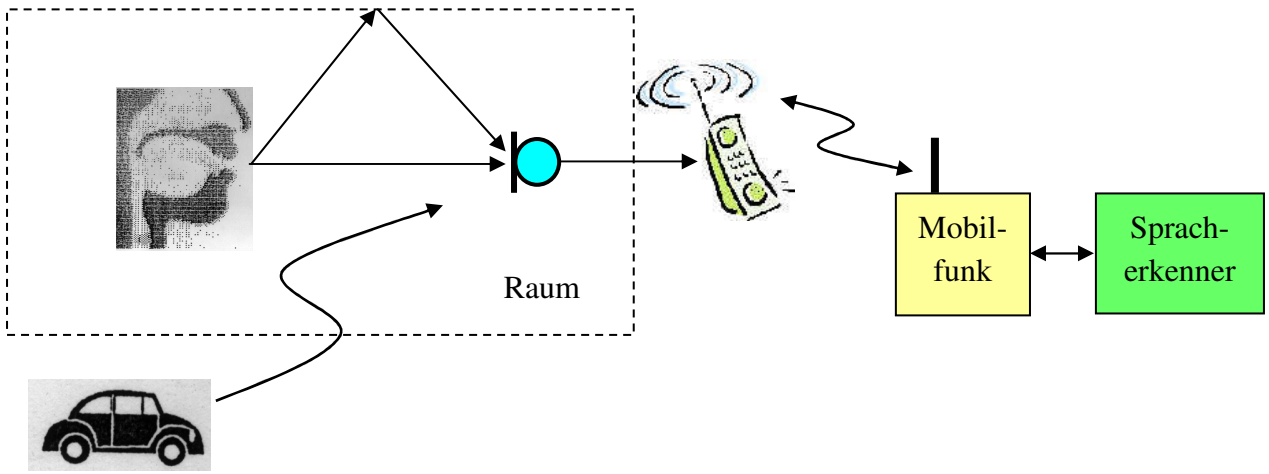

Bild 1: Störeinflüsse beim Einsatz eines Spracherkennungssystems

Die Details dieses Simulationswerkzeugs werden in den Veröffentlichungen [1], [3] und [4] vorgestellt. Damit konnten künstlich gestörte Sprachdaten erzeugt werden, die für die Erkennungsexperimente in den nachfolgenden Phasen des Projekts benötigt wurden.

Im zweiten Abschnitt des Projekts wurden die verschiedenen Störeinflüsse bei der Spracheingabe auf die daraus resultierende Verschlechterung der Erkennungsraten hin untersucht. Damit konnte eine Bewertung der einzelnen Störfaktoren bezüglich ihres Einflusses auf die Erkennungsgüte vorgenommen werden.

In der dritten Phase des Projekts wurde ein neuer Ansatz zur Adaption der Parameter von Hidden Markov Modellen, die als Referenzmuster zur Spracherkennung verwendet werden, auf den Einfluss des Nachhalls bei einer Aufnahme im Freisprechmodus in einer räumlichen Umgebung entwickelt. Zudem wurde die Möglichkeit der Kombination dieser Adaption mit einem bereits vorhandenen Ansatz zur Adaption auf ein sich additiv überlagerndes Störgeräusch sowie auf den Einfluss unbekannter Frequenzgänge auf Grund des Mikrofons oder des Übertragungskanals untersucht.

Im vierten Abschnitt des Projekts wurde der Ansatz der Kombination einer Extraktion robuster akustischer Merkmale mit einer Adaption der Referenzmuster betrachtet. Es wurden verschiedene Möglichkeiten der Kombination auf ihre Eignung zur Verbesserung der Erkennung untersucht.

Die Arbeiten wurden im Wesentlichen in der Form und Reihenfolge durchgeführt, wie sie auch im Antrag aufgeführt wurden. Es kam zu einer zeitlichen Verzögerung in der zweiten Projekthälfte, da der bis dahin in dem Projekt tätige Mitarbeiter, Herr Dr. Finster, zu einem Unternehmen in der freien Wirtschaft wechselte. Die verbleibenden Arbeiten wurden nach einer kurzen Phase der Suche nach einem neuen Mitarbeiter von etwas geringer qualifizierten wissenschaftlichen Mitarbeitern und teilweise in Teilzeittätigkeit durchgeführt, so dass es insgesamt zu einer kostenneutralen Verlängerung des Projekts bis zum 30.06.2008 kam.
Bereits im ersten Antrag waren Mittel zur Beschaffung der deutschen Sprachdatenbasis RVG (Regional Variants of German) beantragt worden, die allerdings erst im Fortsetzungsantrag bewilligt wurden. Daher konnten die im ersten Teilprojekt geplanten, parallelen Untersuchungen mit der deutschen Datenbasis erst in der zweiten Hälfte des Projekts durchgeführt werden.

2.3. Ergebnisse

Mit dem in der ersten Projektphase entwickelten Werkzeug zur Simulation verschiedener akustischer Bedingungen bei der Spracheingabe wurde eine neue Sprachdatenbasis generiert. Dabei wurden die in einer Sammlung von Aufnahmen englischsprachiger Ziffern und Ziffernketten enthaltenen ungestörten Sprachsignale verwendet. Diese Sammlung ist unter der Kurzbezeichnung „TIDigits“ bekannt.

Es wurde eine Aufnahme der Ziffern in der gestörten Umgebung eines Kraftfahrzeugs bzw. in einer räumlichen Büroumgebung und in einer Wohnumgebung simuliert. Die gestörten Daten wurden für die weiteren Untersuchungen benötigt. Um sie auch für andere Forschungsaktivitäten zur Verfügung zu stellen, wurden die Sprachdaten auf zwei DVDs mit einer entsprechenden Dokumentation sowie mit einigen Skripten zur Durchführung von Erkennungsexperimenten mit dem Programmpaket HTK (Hidden Markov Model Toolkit der Universität Cambridge) zusammengestellt. Diese Sprachdatensammlung ist unter der Bezeichnung „Aurora-5“ von der Organisation ELRA (European Language Ressource Association) verfügbar (http://www.elra.org). Damit kann diese Datenbasis zum Vergleich verschiedener Ansätze zur robusten Erkennung von in gestörter und verhallter Umgebung aufgenommenen Sprachsignalen benutzt werden. Eine Beschreibung der Datenbasis und der Erkennungsexperimente ist im Internet unter http://aurora.hs-niederrhein.de verfügbar. Neben der Generierung dieser neuen Sprachdatensammlung wurde unter http://dnt.kr.hs-niederrhein.de/sireac.html eine Benutzer-Schnittstelle im Internet geschaffen, um das Simulationswerkzeug vorzustellen und für jedermann akustisch erfahrbar zu machen. Das Aussehen dieser Schnittstelle wird in Bild 2 veranschaulicht.

Ein Benutzer kann dabei eine bestimmte Situation der Spracheingabe durch Auswahl einer räumlichen Umgebung, durch Wahl eines Störgeräuschs und einer möglichen Sprachcodierung und Funkkanalstörung definieren. Es besteht die Möglichkeit, dass der Benutzer seine eigenen Sprachsignale verwendet und deren Aufnahme entsprechend der eingestellten Konfiguration simuliert wird. Die Ergebnisse aller angestellten Simulationen werden tabellarisch angeboten und erlauben den direkten akustischen Vergleich. Das Simulationswerkzeug und die neue Sprachdatensammlung werden in den Veröffentlichungen [1], [3], [4] und [9] vorgestellt.

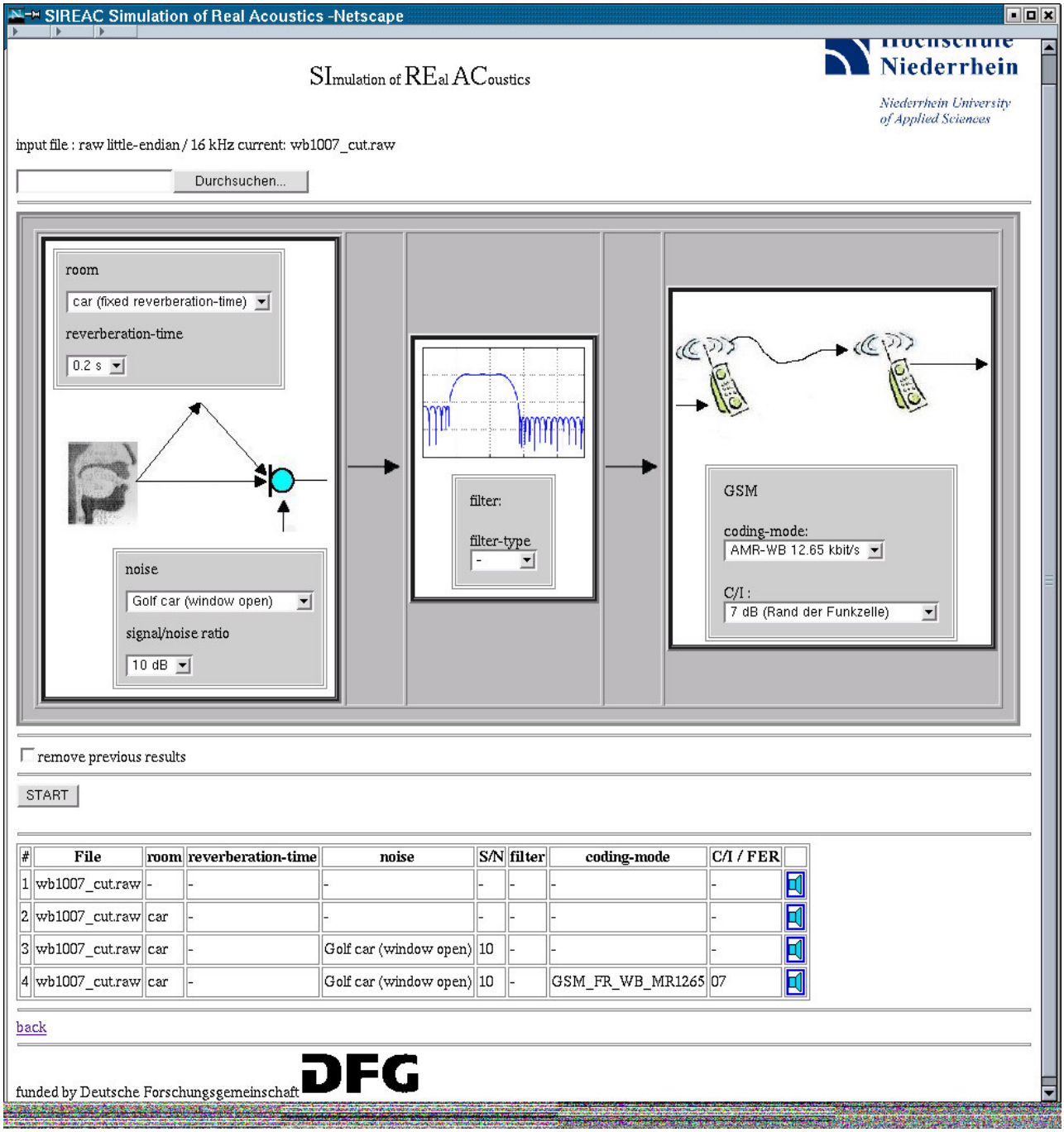

Bild 2: Internetseite zur Präsentation des Simulationswerkzeugs http://dnt.kr.hsnr.de/sireac.html

In der zweiten Phase des Projekts wurde das Simulationswerkzeug eingesetzt, um Sprachdaten unter verschiedenen Aufnahmebedingungen zu erzeugen. Anhand der Verschlechterung der Erkennungsrate bei den unterschiedlichen Aufnahmeszenarien sollte eine quantitative Bewertung der einzelnen Störeinflüsse als auch von Kombinationen von Störeinflüssen vorgenommen werden. Zunächst wurden die Aufnahmen der englischen Ziffern und Ziffernketten der TIDigits Datenbasis als ungestörte Sprachsignale verwendet. Als Ergebnis stellte sich heraus, dass sowohl die additive

Überlagerung von Hintergrundstörungen als auch die Aufnahme im Freisprechmodus in einer verhallten Umgebung die Erkennungsgüte deutlich verschlechtern. Der Einfluss einer Sprachcodierung, wie sie bei einer Übertragung in einem Mobilfunknetz eingesetzt wird, ist vergleichsweise gering. Die Erkennungsrate verschlechtert sich erst dann deutlich, wenn man einen stark gestörten Funkkanal betrachtet. Dies lässt sich jedoch bei einem sinnvollen Einsatz der AMR (Adaptive Multi-Rate) Sprachcodierung, bei der in Abhängigkeit der Qualität des Funkkanals eine leistungsfähigere Kanalcodierung eingesetzt wird, weitgehend vermeiden. Die detaillierten Ergebnisse dieser Projektphase werden in [2] und [3] vorgestellt.

Im dritten Abschnitt dieses Vorhabens wurde ein neuer Ansatz zur Adaption der Referenzmuster auf den akustischen Einfluss einer verhallten, räumlichen Umgebung bei einer Aufnahme im Freisprechmodus untersucht. Die Entwicklung dieses neuen Adaptionsverfahrens stellt den wichtigsten Aspekt dieses Vorhabens dar.
Der Einfluss des Halls eines Raums lässt sich näherungsweise durch eine Impulsantwort beschreiben, die einen exponentiell abfallenden Verlauf über der Zeit besitzt, wie es in Bild 3 dargestellt ist.

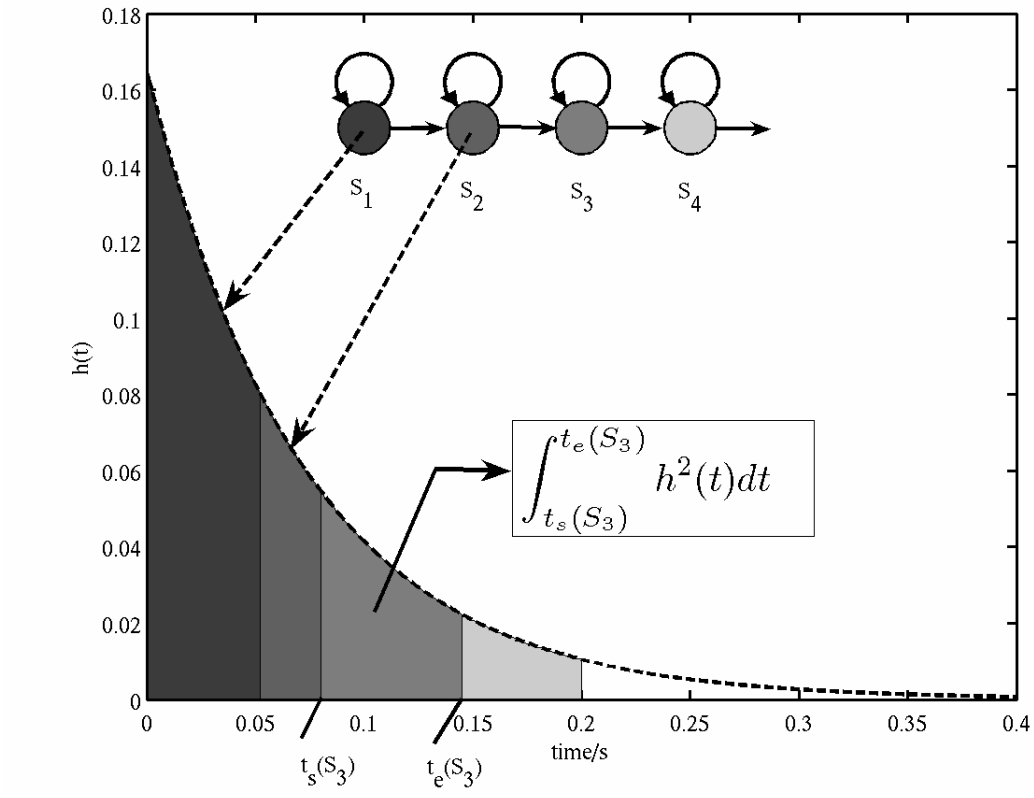

Bild 3: Exponentiell abfallende Impulsantwort zur näherungsweisen Beschreibung des Nachhalls

Die Nachhallzeit ist der Parameter, der den genauen Verlauf des Abfalls definiert. Auf Grund dieser exponentiell abfallenden Impulsantwort treten die spektralen und energetischen Merkmale eines Sprachsignalabschnitts in abgeschwächter Form auch zu späteren Zeitpunkten auf und überlagern sich dem Spektrum des späteren Zeitpunkts. Dieser Ansatz der Überlagerung des Spektrums und der Energie eines Signalabschnitts mit den entsprechend abgeschwächten Merkmalen früherer Abschnitte lässt sich auf die Folge von Zuständen eines HMM übertragen und zur Adaption von HMMs auf eine Spracheingabe in verhallter Umgebung einsetzen.
Beispielhaft werden dazu in Bild 3 die ersten vier Zustände eines HMMs gezeigt. Aus der Übergangswahrscheinlichkeit, die das Verweilen in einem Zustand quantitativ beschreibt, lässt sich die mittlere Dauer des durch den entsprechenden Zustand modellierten Sprachsignalabschnitts bestimmen. Mit Hilfe der Dauer lässt sich der Anteil der Energie, mit dem die charakteristischen Merkmale eines Signalabschnitts zu einem späteren Zeitpunkt auftreten, als Integral über die quadrierte Impulsantwort berechnen, wie es beispielhaft für den dritten Zustand in Bild 3 dargestellt ist. Zur Adaption wird eine Rücktransformation der Mittelwerte der Cepstralkoeffizienten, die das Kurzzeit-Spektrum eines HMM Zustands definieren, in den Bereich des linearen Mel Spektrums vorgenommen. Das Kurzzeit Leistungsdichtespektrum $|X(S_i)|^2$ eines Zustands S_i wird dabei gemäß der nachstehend beschriebenen, additiven gewichteten Überlagerung von Spektren angepasst:

$$\begin{aligned}\left|\tilde{X}_k(S_i)\right|^2 &= \alpha_{i,i}\cdot\left|X_k(S_i)\right|^2 + \alpha_{i,i-1}\cdot\left|X_k(S_{i-1})\right|^2 + \quad \alpha_{i,i-2}\cdot\left|X_k(S_{i-2})\right|^2 + \cdots \\ &= \sum_{j=1}^{i}\alpha_{i,j}\cdot\left|X_k(S_j)\right|^2 \qquad for\ 1\le k\le NR\_mel\end{aligned}$$

Dabei fließen das Spektrum des Zustands i sowie die Spektren aller vorherigen Zustände j=1,2,…,i-1 mit den jeweiligen Wichtungsfaktoren $\alpha_{i,j}$ ein, die den energetischen Anteil der spektralen Merkmale des Zustands j, der zum Zeitpunkt des Zustands i auftritt, festlegen. Der Index k nimmt dabei einen Wert zwischen 1 und der Anzahl NR_mel der Frequenzbänder des Mel Spektrums an.

Die Adaption kann individuell vor jeder neuen Spracheingabe vorgenommen werden. Im einfachsten Fall wird zur Adaption eine Schätzung des Werts der als frequenzunabhängig angenommenen Nachhallzeit benötigt. In diesem Fall besitzen die Wichtungsfaktoren keine Abhängigkeit von der Frequenz. Die Schätzung wird jeweils nach einer Erkennung in Form einer „maximum likelihood“ Bestimmung vorgenommen. Dabei werden für eine geringfügige Variation

der zuvor geschätzten Nachhallzeit jeweils die adaptierten HMMs bestimmt. Es wird die Wahrscheinlichkeit für eine nochmalige Erkennung mit den leicht unterschiedlich adaptierten HMMs berechnet. Die Schätzung der Nachhallzeit wird aus dem Satz adaptierter HMMs abgeleitet, für die die größte Wahrscheinlichkeit berechnet wird.

Neben dem zuvor beschriebenen Verfahren zur Adaption der spektralen Parameter auf den Einfluss des Nachhalls wurde auch ein neues Verfahren zur Adaption der Ableitungen des zeitlichen Verlaufs des Kurzzeit-Spektrums entwickelt. Mit Hilfe dieser Adaption der sogenannten „Delta“ und „Delta-Delta“ Parameter konnte eine weitere Verbesserung der Erkennungsraten erzielt werden. Beispielhaft sind dazu in Bild 5 die Fehlerraten einer Erkennung der TIDigits Sprachdaten in zwei verschiedenen räumlichen Umgebungen dargestellt.

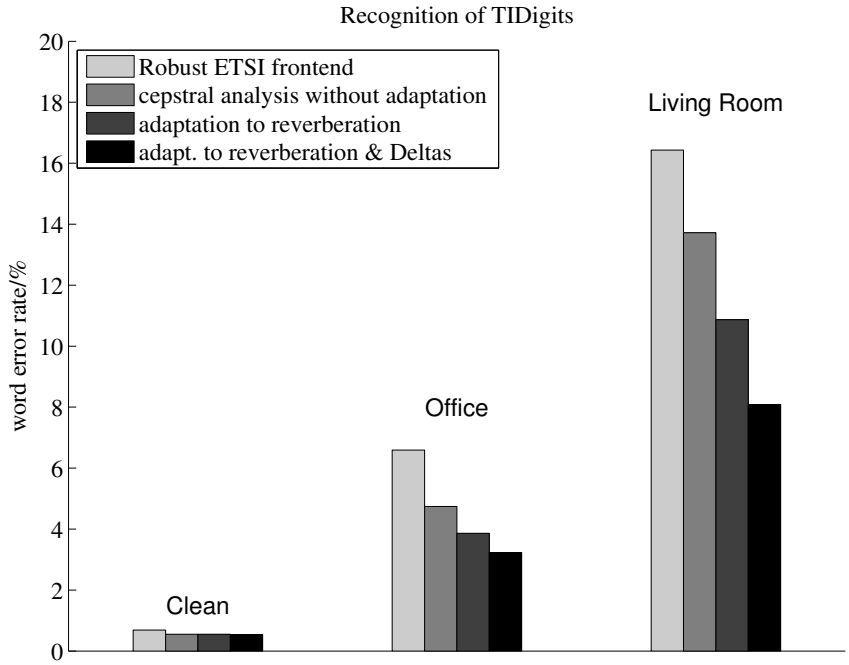

Bild 5: Fehlerraten bei der Erkennung der TIDigits im Freisprechmodus

Man kann eine deutliche Verbesserung der Erkennung bei Einsatz der Halladaption und der Adaption der Delta Parameter gegenüber einem von ETSI standardisierten Verfahren, bei dem robuste akustische Merkmale extrahiert werden, aber keine Adaption bei der Erkennung vorgenommen wird, feststellen.

Im Anschluss konnte des Weiteren gezeigt werden, dass dieses neue Adaptionsverfahren mit einer zuvor bereits entwickelten Adaption auf Hintergrundstörungen und auf unbekannte Frequenzgänge

kombiniert werden kann. Damit lassen sich erhebliche Verbesserungen der Erkennungsraten bei Aufnahmen in einer gestörten räumlichen Umgebung im Freisprechmodus erzielen. Die detaillierten Ergebnisse werden in [1], [2] und [3] vorgestellt. Es konnte ebenfalls gezeigt werden, dass der neue Ansatz auf eine lautbasierte Erkennung größerer Vokabularien angewendet werden kann.

In der letzten Phase des Projekts wurden einige Möglichkeiten untersucht, die Adaption der Referenzmuster mit Verfahren, die im Bereich der Sprachanalyse auf einer Extraktion robuster akustischer Merkmale beruhen, zu verknüpfen. Bisher wird entweder eine Extraktion robuster Merkmale oder eine Adaption eingesetzt. Da der rechnerische Aufwand im Bereich der Merkmalsextraktion in der Regel deutlich niedriger ist, beruhen die meisten in der Praxis eingesetzten Spracherkennungssysteme auf diesem Ansatz.
Als eine Möglichkeit der Kombination wurde ein auf einer Extraktion robuster Merkmale beruhendes Erkennungssystem, dessen Merkmalsextraktion als ETSI Standard definiert ist, um die in der vorherigen Projektphase untersuchte Adaption der HMMs bei einer Spracheingabe im Freisprechmodus erweitert. Da die standardisierte Merkmalsextraktion ebenfalls auf einer Cepstralanalyse beruht, konnte das Adaptionsverfahren nahezu unverändert eingesetzt werden. Es wurde dabei nur eine Adaption auf den Nachhall, jedoch keine Adaption auf Hintergrundstörungen und unbekannte Frequenzgänge vorgenommen, da diese Störeinflüsse bereits durch die Merkmalsextraktion kompensiert werden. Bild 6 veranschaulicht diese Vorgehensweise.

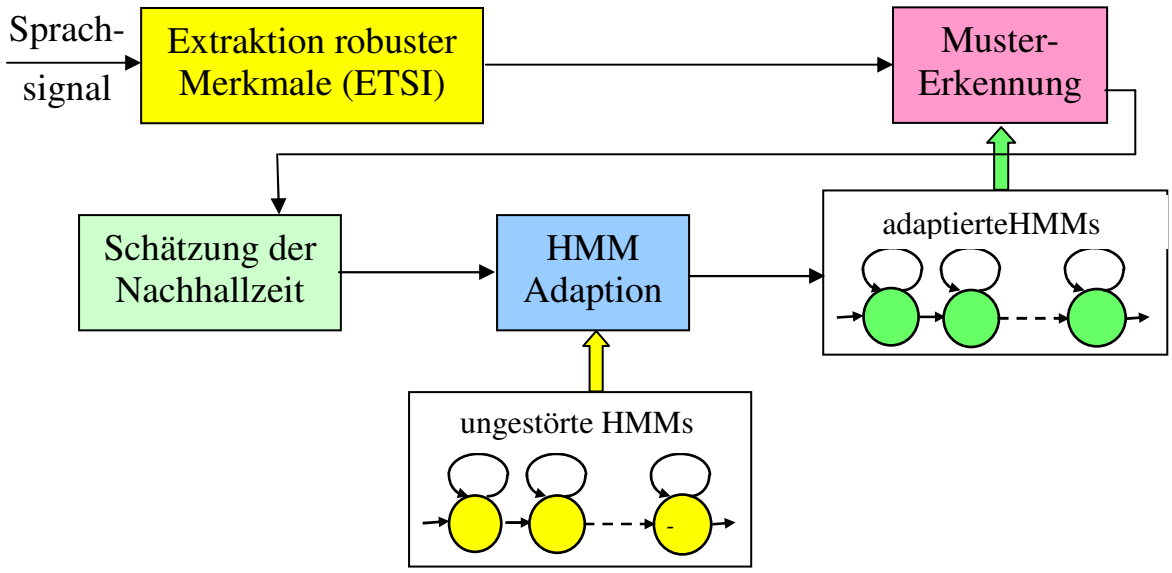

Bild 6: Kombination einer robusten Merkmalsextraktion und einer Adaption der HMMs auf Nachhall

In Bild 7 werden die Fehlerraten dargestellt, die sich für zwei der in der Datenbasis Aurora-5 betrachteten Störbedingungen einstellen. Es werden die Aufnahmen bei alleinigem Auftreten von Hintergrundstörungen sowie im Freisprechmodus in einer Wohnzimmerumgebung, in der die gleichen Störgeräusche auftreten, betrachtet.

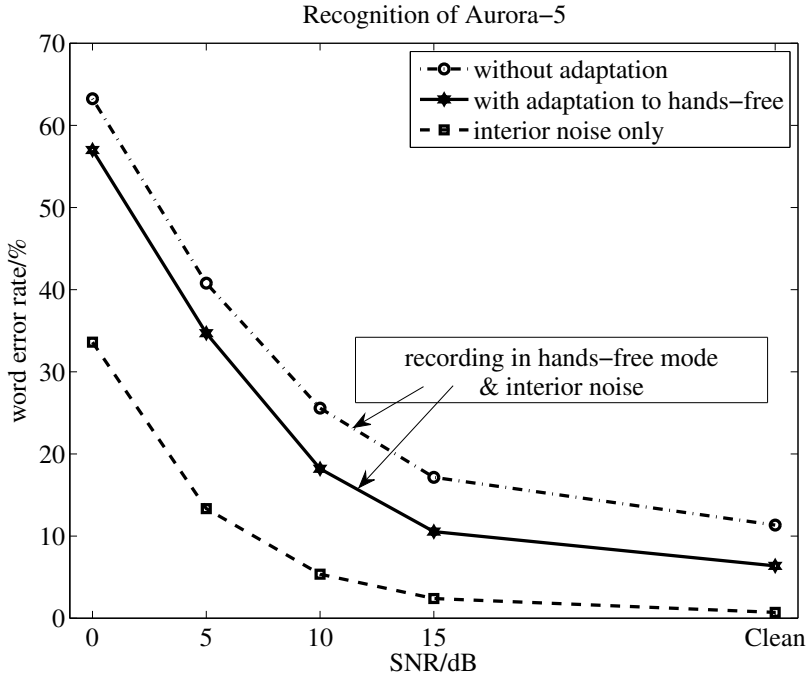

Bild 7: Fehlerraten bei einer Aufnahme gestörter Daten im Freisprechmodus

Zunächst kann man eine deutliche Verschlechterung der Erkennung feststellen, wenn neben dem Auftreten der Störgeräusche zusätzlich die Aufnahme im Freisprechmodus in der räumlichen Umgebung erlaubt wird. Durch die Anwendung des Adaptionsverfahrens ist man in der Lage die Fehlerraten deutlich zu reduzieren. Das entwickelte Adaptionsverfahren lässt sich gut mit dem standardisierten Verfahren zur Extraktion robuster Merkmale kombinieren. Die erzielten Ergebnisse werden in [6] veröffentlicht werden.

Neben der von ETSI standardisierten Merkmalsextraktion wurden alternativ einige Verfahren zur einkanaligen Störreduktion auf ihre Verwendbarkeit zur Verbesserung der Erkennungsraten untersucht. Diese Ansätze werden am Institut für Kommunikationsakustik an der Ruhr Universität entwickelt mit der Zielsetzung einer Sprachverbesserung in gestörten Kommunikationssituationen. Die Verfahren beruhen auf einer adaptiven Filterung, die auf das mit Hilfe der DFT ermittelte Kurzzeit-Spektrum angewendet wird. Damit lassen sich diese Ansätze in die meisten im Bereich der Spracherkennung eingesetzten Verfahren zur Extraktion akustischer Merkmale mit relativ geringem

Aufwand integrieren. Dabei konnten bei der Erkennung gestörter Sprachsignale vergleichbare Erkennungsergebnisse wie mit dem von ETSI standardisierten Verfahren erzielt werden. Es wurden jedoch keine darüber hinausgehenden Verbesserungen erzielt.

Die meisten der im diesem Projekt angestellten Erkennungsexperimente beruhten auf einer Erkennung gestörter Versionen der englischsprachigen TIDigits Sprachdaten. Daneben wurde in der zweiten Projekthälfte die deutschsprachige Datenbasis, die unter der Kurzbezeichnung RVG (Regional Variants of German) geführt wird, beschafft. Damit wurde das Ziel verfolgt, die Übertragbarkeit der erzielten Ergebnisse auf andere Erkennungsaufgaben in einer anderen Sprache nachweisen zu können. Zunächst ließ sich feststellen, dass viele der in RVG enthaltenen Aufnahmen bereits bei Vorhandensein von Störgeräuschen im Hintergrund erstellt wurden. Damit musste eine Auswahl von Aufnahmen getroffen werden, bei denen keine Störungen vorhanden sind, um damit das Testmaterial für definierte Störbedingungen erzeugen zu können. Prinzipiell wurden bei den auf den RVG Daten basierenden Experimenten die gleichen Verbesserungen zur Erkennung der im Freisprechmodus in einer gestörten Umgebung aufgenommenen Sprachdaten festgestellt, die sich mit den im Rahmen des Projekts entwickelten Verfahren bei die TIDigits erzielen lassen. Die Ergebnisse werden noch in zukünftigen Veröffentlichungen präsentiert werden. Neben den TIDigits und den RVG Daten wurden noch weitere Experimente mit real in gestörten Räumen aufgenommenen Sprachdaten gemacht. Es konnte auch dabei der Nachweis erbracht, dass die entwickelten Verfahren zur gleichen deutlichen Verbesserung der Erkennungsraten wie bei den Experimenten mit künstlich gestörten Daten führen.

2.4. Verwertung

Die im Rahmen des Projekts entwickelten Adaptionsverfahren können prinzipiell bei nahezu allen Spracherkennungssystemen eingesetzt werden. Der zusätzliche rechnerische Aufwand kann als überschaubar angesehen werden, so dass die Integration und Implementierung in einem Echtzeitsystem realisierbar ist. Damit lässt sich der Anwendungsbereich eines Erkennungssystems um eine Spracheingabe im Freisprechmodus in einer gestörten Umgebung erweitern.
Die gewonnenen Erkenntnisse werden in ein zukünftiges, vom BMBF im Rahmen der Förderlinie FHProfUnd gefördertes Projekt des Antragstellers einfließen, bei dem in Kooperation mit einem industriellen Partner und weiteren Partnern aus dem wissenschaftlichen Umfeld die Robustheit von Spracherkennungssystemen weiter verbessert werden soll.

2.5. Mitarbeiter und Kooperationen

In der Zeit vom 01.04.2004 bis zum 31.12.2005 hat Herr Dr.-Ing. Harald Finster als wissenschaftlicher Mitarbeiter das Projekt bearbeitet. Ab dem 01.05.2006 bis zum 29.02.2007 hat Herr Dipl-Ing. Patrick Pogscheba im Rahmen einer Teilzeittätigkeit als wissenschaftlicher Mitarbeiter einige der zuvor beschriebenen Untersuchungen angestellt. In der Zeit vom 01.09.2007 bis zum 30.06.2008 bearbeitete Herr Dipl.-Ing. Andreas Kitzig weitere Teilaspekte des Projekts. Herr Pogscheba und Herr Kitzig hatten zuvor jeweils ihre Diplomarbeiten im thematischen Umfeld des Projekts durchgeführt. Eine Vielzahl weiterer Diplom- und Projektarbeiten haben Beiträge zu einigen der erzielten Ergebnisse geliefert.

Der Antragsteller hat in dem Zeitraum von Mai bis August 2005 während eines gewährten Forschungsfreisemesters einen Forschungsaufenthalt am International Computer Science Institute in Berkeley, USA durchgeführt, während dessen er einige der entscheidenden Entwicklungen und Untersuchungen im wissenschaftlichen Dialog mit den Mitarbeitern der an dem Institut tätigen Sprachverarbeitungsgruppe vornehmen konnte.

Im Rahmen der Untersuchungen zur Entwicklung des Simulationswerkzeugs als auch zum Einsatz von Störunterdrückungsverfahren fand eine relativ intensiver Dialog und eine Kooperation mit Mitarbeitern des Instituts für Kommunikationsakustik an der Ruhr-Universität Bochum (Prof. Rainer Martin) und des Instituts für Nachrichtengeräte und Datenverarbeitung an der RWTH Aachen (Prof. Peter Vary) statt. Ebenfalls fand ein Dialog und Austausch mit Mitarbeitern des Lehrstuhls für Multi-Media Kommunikation und Signalverarbeitung an der Universität Erlangen (Prof. Walter Kellermann) statt, die an alternativen Ansätzen zur robusten Erkennung verhallter Sprache arbeiten.

2.6. Qualifizierung

Herr Dr. Finster hat seine Tätigkeit in dem Projekt beendet, da er auf Grund seiner erworbenen Kenntnisse im Bereich der robusten Spracherkennung ein attraktives Angebot zur Mitarbeit in der Firma Topsystems in Würselen erhalten hatte, die einen Spezialisten für das Themengebiet der Spracherkennung suchte.

Herr Pogscheba und Herr Kitzig haben auf Grund ihres durch die Mitarbeit in dem Projekt geweckten Interesses an der wissenschaftlichen Arbeit beide noch weitergehende Master-Studien aufgenommen. Daneben hat durch eine Vielzahl von Diplom- und Projektarbeiten eine weitere Qualifizierung des wissenschaftlichen Nachwuchses stattgefunden.

3. Zusammenfassung

Es wurde ein Verfahren zur automatischen Spracherkennung entwickelt, mit dem die Erkennungsraten bei einer Spracheingabe im Freisprechmodus in einer gestörten räumlichen Umgebung im Vergleich zu bisherigen Verfahren verbessert werden können. Das Verfahren beruht auf einer Adaption der zur Erkennung verwendeten Referenzmuster, die als Hidden Markov Modelle (HMMs) zur Mustererkennung herangezogen werden. Speziell werden die in den HMMs enthaltenen spektralen und energetischen Parameter auf die Hintergrundstörung und den Nachhall des Raumes bei jeder neuen Spracheingabe adaptiert. Neu ist dabei die Adaption auf den Nachhall, der auf Grund der Vielfachreflexionen des Schalls in einem Raum zu einem zeitlich ausgedehnten Auftreten der spektralen und energetischen Merkmale der Sprache führt. Dies wird bei der Adaption der Merkmale eines HMM Zustands durch eine additive, gemäß dem jeweiligen Nachhallverhalten gewichtete Überlagerung der Merkmale vorheriger Zustände kompensiert. Des Weiteren wurde ein neues Verfahren zur Adaption der zeitlichen Ableitungen der akustischen Parameter, die häufig als Delta und Delta-Delta Parameter bezeichnet werden, entwickelt.

Zur Durchführung von Erkennungsexperimenten wurde ein Simulationswerkzeug entwickelt, um die Aufnahme in einer gestörten und verhallten Umgebung sowie eine mögliche Übertragung über einen Mobilfunkkanal nachzuempfinden. Damit wurden Versionen der ungestörten TIDigits Sprachdatenbasis erzeugt, die die zu untersuchenden Aufnahmebedingungen beinhalten. Diese Sammlung von Sprachdaten steht allen Forschungsgruppen als so bezeichnete „Aurora-5“ Datenbasis zur Verfügung, die von der für die Verteilung von Sprachdaten zuständige Organisation ELRA bezogen werden kann. Mit Hilfe von Erkennungsexperimenten mit der neu geschaffenen Sprachdatenbasis als auch mit real in Räumen aufgenommenen Sprachdaten konnte die Effizienz der neuen Verfahren zur Verbesserung der Erkennungsraten aufgezeigt werden.

Abschließend konnte ebenfalls gezeigt werden, dass durch eine Kombination bekannter Erkennungsverfahren, die auf einer Extraktion robuster akustischer Merkmale beruhen, und der Adaption der Referenzmuster auf den Nachhall eines Raumes, die Leistungsfähigkeit der bestehenden Verfahren bei einer Spracheingabe im Freisprechmodus verbessert werden kann.