Hochschule Niederrhein
University of Applied Sciences

Ergebnisse des 2. Projektabschnitts

Das Ziel des zweiten Projektabschnitts ist eine Auswahl von Sprachdatensammlungen und eine Definition von Spracherkennungsexperimenten, um die im ersten Abschnitt festgelegten akustischen Szenarien untersuchen zu können. Zudem werden erste Erkennungsergebnisse für die definierten Experimente erzeugt, wobei zur softwaremäßigen Realisierung der Experimente sowohl standardisierte als auch eigene Verfahren zur Extraktion der akustischen Merkmale und zur Mustererkennung mit und ohne Adaption herangezogen werden.

Im ersten Projektabschnitt hatten sich die akustischen Szenarien

- des Freisprechens in einem fahrenden Kraftfahrzeug und
- des Freisprechens in einer störschallerfüllten, räumlichen Umgebung

als die beiden im Rahmen dieses Projekts zu untersuchenden Bedingungen herausgestellt. Bevor die für die jeweilige akustische Umgebung ausgewählten Sprachdatensammlungen sowie die definierten Erkennungsexperimente vorgestellt werden, werden die für die Durchführung der Untersuchungen allgemein verwendeten Verfahren und Werkzeuge zur Spracherkennung erläutert.

Eingesetzte Verfahren zur Spracherkennung

Ein Spracherkennungssystem besitzt den in Bild 1 dargestellten Aufbau.

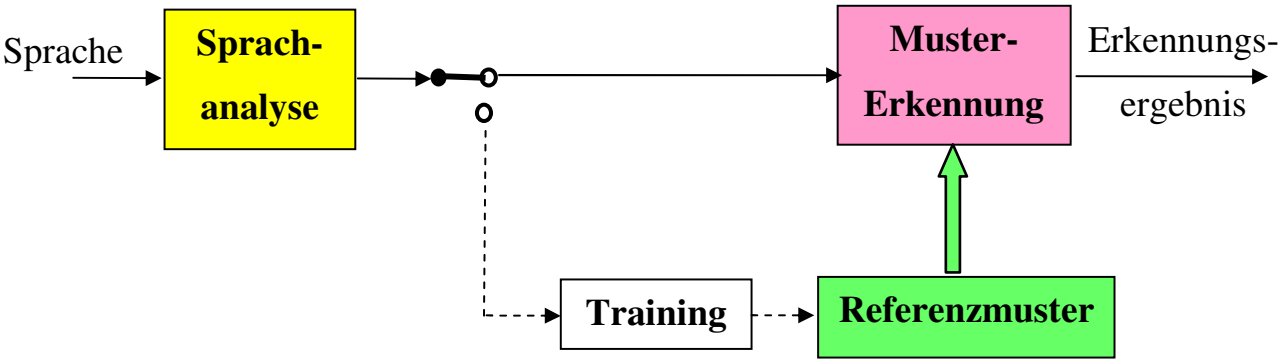

Bild 1 : Allgemeiner Aufbau eines Spracherkennungssystems

Zunächst werden dem Sprachsignal durch eine Analyse kurzer Signalabschnitte verschiedene akustische Parameter entnommen, die die charakteristischen Merkmale zur Unterscheidung der unterschiedlichen Laute einer Sprache beinhalten. Dabei extrahieren die meisten heutzutage

eingesetzten Verfahren die sogenannten Cepstral-Koeffizienten, die das Kurzzeitspektrum des analysierten Signalabschnitts beinhalten, sowie einen Energieparameter. Ergänzt werden diese Parameter um die ersten und zweiten Ableitungen des zeitlichen Verlaufs jedes Merkmalswerts, die man als Delta bzw. Delta-Delta Koeffizienten bezeichnet. Im Rahmen dieses Projekts werden die beiden nachstehenden Verfahren eingesetzt:

1. Im Rahmen vorhergehender Arbeiten wurde vom Autor ein auf einer DFT beruhendes Verfahren zur Extraktion von 12 oder 13 cepstralen Koeffizienten und eines logarithmierten Energie-Koeffizienten aus den 25 ms langen Abschnitten eines mit 8 kHz abgetasteten Signal entwickelt. Die insgesamt 13 oder 14 akustischen Parameter werden um die Delta und Delta-Delta Koeffizienten erweitert, so dass ein Merkmalsvektor insgesamt 39 bzw. 42 akustische Parameter beinhaltet. Die Analyse findet alle 10 ms statt. Zur Referenzierung dieses Verfahrens wird nachfolgend das Kürzel „HGH“ verwendet,
2. Von ETSI, einem Institut, dessen Aufgabe die Standardisierung der in den heutigen Mobilfunknetzen eingesetzten Verfahren und Protokolle ist, wurde ein Verfahren zur Extraktion robuster akustischer Merkmale festgelegt. Ähnlich wie bei dem zuvor angeführten Verfahren werden 13 cepstrale Koeffizienten und 1 Energieparameter extrahiert. Dabei wird in der Signalanalyse ein auf einer Wiener Filterung beruhendes Verfahren zur Unterdrückung stationärer Hintergrundstörungen eingesetzt. Zudem wurde ein Verarbeitungsblock zur Schätzung und Kompensation unbekannter Frequenzgänge, wie sie beispielsweise aus dem Einsatz unterschiedlicher Mikrofontypen resultieren, integriert. Mit diesem Verfahren, das nachstehend mit dem Kürzel „ETSI-2“ referenziert wird, können beachtliche Verbesserungen der Erkennungsraten bei der Erkennung gestörter Sprachsignale erzielt werden. Es kann als eines der zurzeit besten Verfahren, das auf einer Extraktion robuster Merkmale beruht, angesehen werden.

Für beide Verfahren sind Implementierungen in C und Matlab vorhanden, wobei der C Code für das zweite Verfahren von ETSI zur Verfügung gestellt wird.

Als Referenzmuster werden die in den meisten Spracherkennungsverfahren eingesetzten Hidden-Markov Modelle (HMM) verwendet. Zum Training der Modelle werden die Programme des Hidden-Markov Model Toolkits (HTK), das an der Universität von Cambridge entwickelt wurde, eingesetzt. HTK beinhaltet eine frei verfügbare Sammlung von Programmen zur Durchführung kompletter Spracherkennungsexperimente. Die eigenen Programme wurden so gestaltet, dass sie

die in HTK definierten Fileformate, z.B. zum Speichern der akustischen Merkmale oder der Parameter eines HMMS, unterstützen.
Zur eigentlichen Mustererkennung wurde im Rahmen vorhergehender Arbeiten ein auf dem Viterbi Algorithmus beruhender Ansatz zur Bestimmung der Wahrscheinlichkeiten, dass eine analysierte Folge von Merkmalsvektoren von einem HMM erzeugt werden kann, in C und in Matlab implementiert. Die damit erzielten Erkennungsergebnisse sind identisch mit denjenigen, die mit dem in HTK vorhandenen Erkennungsprogramm erzielt werden können. Zudem wurde in einem früheren Projekt ein eigener Ansatz zur Adaption der ein HMM definierenden Parameter an eine unbekannte akustische Umgebung entwickelt und softwaremäßig in C und Matlab realisiert. Die Adaption beruht dabei auf einer Schätzung des Spektrums der Hintergrundstörung, einer Schätzung eines unbekannten Frequenzgangs sowie einer Schätzung der Nachhallzeit. Das Adaptionsverfahren ist auf das Analyseverfahren „HGH“ abgestimmt. Der Ansatz einer Adaption der in den Referenzmustern enthaltenen Merkmale stellt eine Alternative zur Extraktion robuster Merkmale dar, um die Erkennungsleistung in einer gestörten akustischen Umgebung zu verbessern.

Daher werden im Folgenden vergleichend die Erkennungsergebnisse für eine Erkennung mit dem

- „HGH“ Analyseverfahren als Beispiel einer Analyse cepstraler und energetischer Parameter, wie sie in den meisten heutzutage verwendeten Erkennungssystemen eingesetzt wird,
- „ETSI-2“ Analyseverfahren als Beispiel eines auf einer robusten Merkmalsextraktion beruhenden Erkennungssystems,
- „HGH“ Analyseverfahren in Kombination mit einer Adaption der Referenzmuster

betrachtet. Der erstgenannte Ansatz wird zur Bestimmung von Ergebnissen herangezogen, die mit einem heutigen System erzielt werden, in dem keine zusätzlichen Maßnahmen zur Erhöhung der Robustheit getroffen werden. Die beiden weiteren Ansätze dienen als Beispiele von Systemen, die auf einer der beiden alternativen Vorgehensweisen zur Verbesserung der Erkennung in gestörten Situationen beruhen.

Freisprechen in einem fahrenden Kraftfahrzeug

Als Erkennungsaufgaben werden die Erkennung von englischen und von deutschen Ziffernketten betrachtet. Zur Erkennung englischer Ziffernketten wurden vom Autor im Rahmen früherer Arbeiten bereits verschiedene Experimente zur Erkennung der Ziffern in gestörten akustischen

Umgebungen definiert. Die zugehörigen Sprachdaten und entsprechende Software Werkzeuge wurden der Allgemeinheit zur Durchführung vergleichender Experimente zur Verfügung gestellt und sind bei der Organisation mit dem Namen ELRA (European Language Resource Association) unter der Bezeichnung „Aurora-5“ erhältlich. Die Basis der Aurora-5 Daten bilden die unter der Bezeichnung TIDigits bekannten Aufnahmen der englischen Ziffern, die bei der Firma Texas Instruments von amerikanischen Sprechern aufgezeichnet wurden und der Allgemeinheit zur Verfügung gestellt wurden. Die TIDigits wurden in sehr ruhiger Umgebung aufgenommen. Daher sind sie gut geeignet, durch die additive Überlagerung von Hintergrundstörungen oder die Simulation eines Freisprechens in einer räumlichen Umgebung verschiedene Versionen dieser Sprachdaten zu erzeugen, mit denen man die Aufnahme in unterschiedlichen akustischen Umgebungen nachempfinden kann. In „Aurora-5“ gibt es Versionen der zur Erkennung vorgesehenen Aufnahmen, bei denen jeweils ein Abschnitt der in einem fahrenden Auto aufgenommenen Hintergrundstörungen bei einem definierten Signal-/Störleistungsverhältnis (SNR) überlagert wurde. Zudem wird neben dem Störgeräusch noch das Freisprechen im Kraftfahrzeug berücksichtigt.

Als Referenzmuster werden aus den ungestörten, für das Training vorgesehenen Sprachdaten geschlechtsspezifische HMMs generiert. Die HMMs besitzen 16 Zustände. Jeder akustische Parameter wird in jedem Zustand durch eine Mischverteilung zweier Gauß-Verteilungen modelliert. Als Pausenmodell wird ein Modell mit einem Zustand und ? Gauß-Verteilungen verwendet. Tabelle 1 beinhaltet die Wortfehlerraten für die verschiedenen Analyseverfahren sowie für eine Erkennung mit Adaption. Jede Wortfehlerrate resultiert aus der Wortketten-Erkennung von 8700 Äußerungen mit 28583 gesprochenen Ziffern. Die gleichen Ergebnisse werden in Bild 2 graphisch veranschaulicht.

		SNR/dB				
Analyseverfahren	Adaption	Clean	15	10	5	0
ETSI-2		0,55 %	4,17 %	7,62 %	15,42 %	33,48 %
HGH		0,55 %	14,64 %	35,25 %	65,04 %	84,48 %
HGH	X	0,53 %	2,09 %	4,93 %	15,26 %	41,74 %

Tabelle 1: Wortfehlerraten für „Aurora-5“ in der akustischen Umgebung „car noise handsfree“

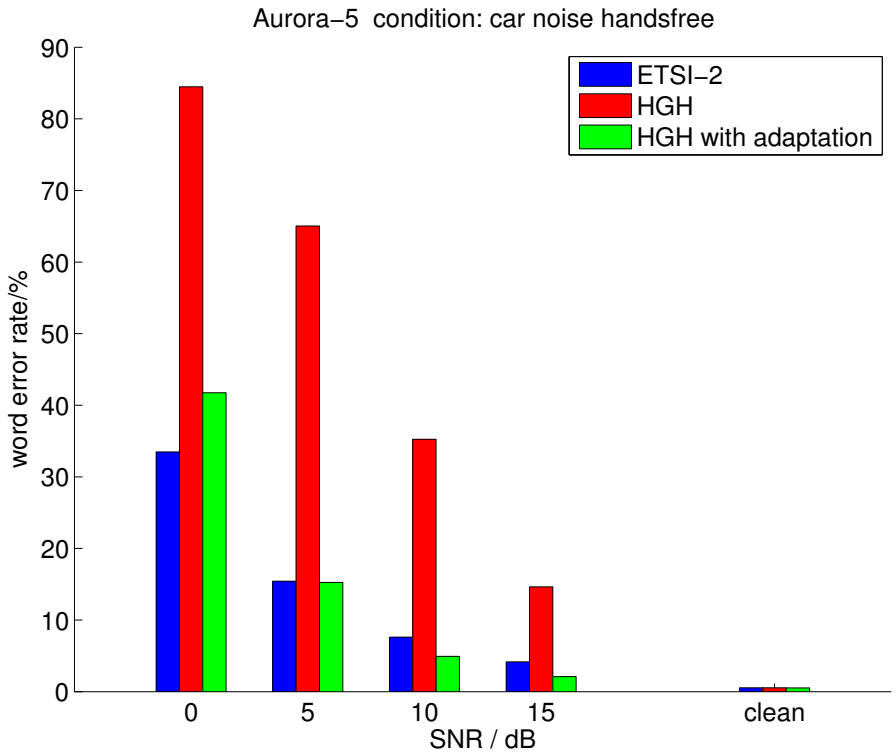

Bild 2 : Wortfehlerraten für „Aurora-5“ bei Aufnahme in Autos im Freisprechmodus

Die Ergebnisse, wie sie der mittleren Zeile in Tabelle 1 bzw. den roten Balken in Bild 2 entnommen werden können, verdeutlichen die Verschlechterung, die sich bei einem Spracherkennungssystem einstellt, bei dem keine Maßnahmen zur Kompensation des Einflusses stationärer Hintergrundstörungen getroffen werden. Der Vergleich mit den beiden alternativ untersuchten Verfahren, die auf einer Verwendung robuster akustischer Merkmale bzw. einer Adaption der Referenzmuster beruhen, zeigt, dass die Fehlerraten durch die zusätzlichen Verarbeitungsschritte deutlich reduziert werden können.

Zur Erkennung deutscher Ziffernketten wird als Basis die Sprachdatensammlung mit der Bezeichnung „RVG“ (Regional Variants of German) verwendet, die am Institut für Phonetik an der TU München erstellt wurde und die von ELRA bezogen werden kann. Die Aufnahmen der RVG Datensammlung besitzen eine deutlich schlechtere Qualität als die TIDigits, da die Qualität der zur Aufzeichnung verwendeten Geräteanordnung schlechter war und die Aufnahmen teilweise im Vorhandensein von Hintergrundstörungen erstellt wurden. Daher wurden bei der Zusammenstellung der zur Erkennung verwendeten Sprachdaten nur Aufnahmen ausgewählt, die ein SNR größer als 25 dB besitzen. Das SNR wird dabei mit einem eigenen Software-Werkzeug bestimmt. Zudem beinhaltet die „RVG“ Sammlung, wie die Bezeichnung „regional variants“ auch

andeutet, Sprecher verschiedener Regionen mit teilweise recht ausgeprägtem Dialekt. Da die durch dialektale Färbung bei der Aussprache hervorgerufenen Probleme im Rahmen dieser Untersuchungen nicht betrachtet werden, wurden die Sprecher zweier Regionen ausgewählt, bei denen nur eine geringfügige Veränderung der Aussprache auf Grund des Dialekts auftritt. Insgesamt wurden 4767 Aufnahmen ausgewählt, die insgesamt 19947 Ziffern beinhalten. Auf der Basis dieser Aufnahmen wurden gestörte Versionen erzeugt. Dazu wird jeweils ein Abschnitt der in einem fahrenden Auto aufgenommenen Hintergrundstörungen bei einem definierten Signal-/Störleistungsverhältnis überlagert. Zudem wird neben dem Störgeräusch noch das Freisprechen im Kraftfahrzeug berücksichtigt. Das Freisprechen wird durch die Faltung des Sprachsignals mit einer in einem Auto bestimmten Impulsantwort realisiert. Die Impulsantwort wurde aktuell im Rahmen dieses Projekts erzeugt. Die Vorgehensweise zur Bestimmung der Impulsantwort wird im Anhang erläutert.

Als Referenzmuster werden wie bei „Aurora-5“ aus den ungestörten Sprachdaten geschlechtsspezifische HMMs mit 16 Zuständen generiert. Tabelle 2 beinhaltet die Wortfehlerraten für die verschiedenen Analyseverfahren sowie für eine Erkennung mit Adaption. Jede Wortfehlerrate resultiert aus der Wortketten-Erkennung der 4767 Äußerungen mit 19947 gesprochenen Ziffern. Die gleichen Ergebnisse werden in Bild 3 graphisch veranschaulicht.

		SNR/dB				
Analyseverfahren	Adaption	Clean	15	10	5	0
ETSI-2		3,18 %	6,02 %	9,96 %	16,41 %	31,84 %
HGH		3,33 %	10,28 %	17,88 %	33,74 %	59,27 %
HGH	X	3,43 %	6,22 %	8,97 %	15,99 %	30,77 %

Tabelle 2: Wortfehlerraten für „RVG“ in der akustischen Umgebung „car noise handsfree“

Zunächst fällt bei der Betrachtung der Ergebnisse auf, dass die Wortfehlerraten bei der Erkennung der „ungestörten“ deutschen Daten höher sind im Vergleich zur Erkennung der englischen Ziffernketten. Dies liegt darin begründet, dass bei den Aufnahmen der RVG Sammlung nicht durchgehend eine hochwertige Aufnahmeanordnung verwendet wurde und auf eine ruhige Aufnahmeumgebung geachtet wurde, so dass die ungestörten Daten bereits ein schlechteres SNR

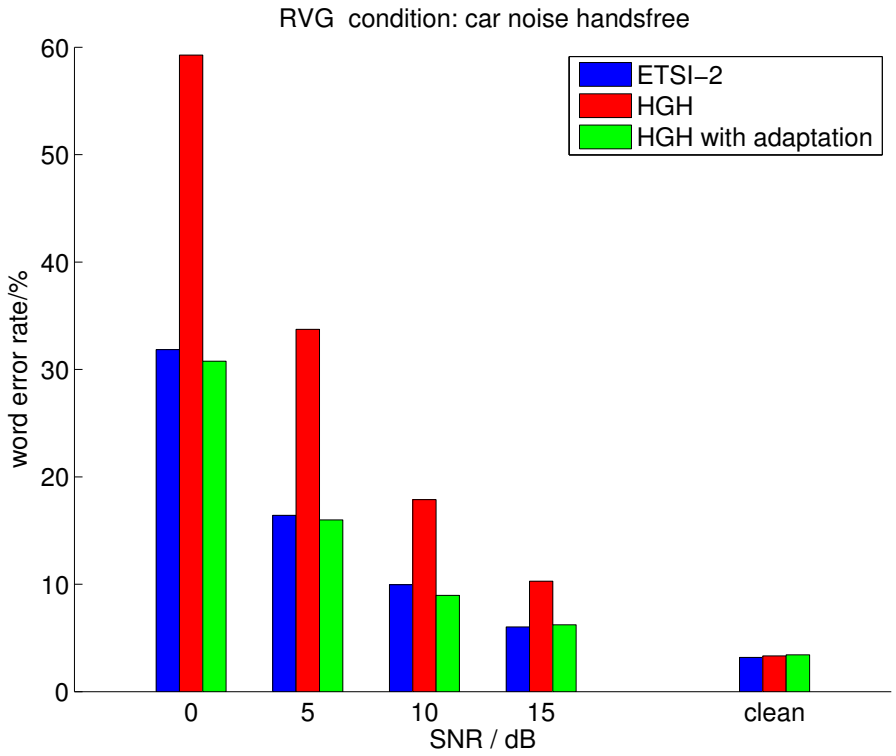

Bild 3: Wortfehlerraten für „RVG“ bei Aufnahme in Autos im Freisprechmodus

besitzen. Bei den gestörten deutschen Daten lassen sich prinzipiell die gleichen Effekte wie bei den englischen Daten beobachten.

Neben der Erkennung der künstlich gestörten Sprachsignale der RVG Datenbasis wird die Erkennung von real in Kraftfahrzeugen aufgenommenen Signalen untersucht. Dazu werden deutsche Ziffernketten, die im Rahmen des Projekts „SpeechDat-Car“ aufgezeichnet wurden, herangezogen. Diese Aufnahmen sind ebenfalls Bestandteil der unter der Bezeichnung „Aurora-3“ geführten Sprachdatensammlung. Es stehen die gleichzeitig mit 2 Mikrofonen aufgenommenen Sprachsignale zur Verfügung. Damit können die Erkennungsergebnisse für eine Nahbesprechung und eine Aufnahme im Freisprechmodus bestimmt und miteinander verglichen werden. Es stehen jeweils 1485 Aufnahmen mit insgesamt 8341 gesprochenen deutschen Ziffern zur Verfügung. Die Aufnahmen wurden in verschiedenen Fahrsituationen mit unterschiedlichen Pegeln der Hintergrundstörung erstellt. Als Referenzmuster werden die zuvor erwähnten, aus den ungestörten Aufnahmen der RVG Sammlung bestimmten HMMs verwendet. Die Wortfehlerraten sind in Tabelle 3 zusammengestellt.

Analyseverfahren	Adaption	Nahbesprechung	Freisprechen
ETSI-2		5,48 %	10,24 %
HGH		10,69 %	38,32 %
HGH	X	5,19 %	13,19 %

Tabelle 3: Wortfehlerraten für „SpeechDat-Car“

Die Verwendung robuster akustischer Merkmale und die Adaption der Referenzmuster führen auch bei diesen real in der Störsituation aufgenommenen Sprachsignalen zu einer deutlichen Reduktion der Fehlerraten. Bei den im Freisprechmodus aufgenommenen Signalen ist das SNR deutlich geringer im Vergleich zu den Aufnahmen bei Nahbesprechung, was den Hauptgrund für die höheren Fehlerraten darstellt.

Freisprechen in einer räumlichen Umgebung

Zur Untersuchung des Einflusses des Freisprechens in einer gestörten räumlichen Umgebung wird neben der Erkennung der englischen und deutschen Ziffernketten, die auch für die Experimente zum Freisprechen im Auto verwendet werden, die Aufgabe der Erkennung isoliert gesprochener italienischer Kommandowörter betrachtet.

Bei der Erstellung der „Aurora-5“ Datenbasis wurden Versionen der TIDigits erzeugt, bei denen die Aufnahme in einem wohnzimmergroßen Raum im Freisprechmodus simuliert wird und den verhallten Sprachsignalen ein in dieser Umgebung typischerweise auftretendes Störsignal bei einem definierten Signal-/Störleistungsverhältnis (SNR) überlagert wird. Zur Simulation der Aufnahme im Raum wurde eine vom Institut für Kommunikationsakustik an der Ruhr-Uni Bochum zur Verfügung gestellte Impulsantwort verwendet, die mit einem Werkzeug zur künstlichen Erzeugung derartiger Impulsantworten generiert wurde. Die überlagerten Abschnitte der Störgeräusche wurden Aufnahmen der typischerweise in räumlichen Umgebungen auftretenden akustischen Szenarien entnommen. Im Vergleich zu den in einem Auto auftretenden Hintergrundstörungen besitzen diese Signale in vielen Abschnitten ein nicht stationäres Verhalten. Beispielsweise tritt Musik oder auch Sprache häufiger als Störung im Hintergrund auf. Da die

Erkennungsergebnisse aufgrund dessen auch generell schlechter sind, wird im Folgenden ein Wert von 5 dB als geringstes SNR betrachtet.

Als Referenzmuster werden die gleichen, geschlechtsspezifischen HMMs wie bei den Experimenten zur Erkennung im Auto verwendet, die aus den ungestörten, für das Training vorgesehenen Sprachdaten generiert wurden. Tabelle 4 beinhaltet die Wortfehlerraten für die verschiedenen Analyseverfahren sowie für die Erkennung mit Adaption. Jede Wortfehlerrate resultiert aus der Wortketten-Erkennung von 8700 Äußerungen mit 28583 gesprochenen Ziffern. Die gleichen Ergebnisse werden in Bild 4 graphisch veranschaulicht.

		SNR/dB			
Analyseverfahren	Adaption	Clean	15	10	5
ETSI-2		0,55 %	9,84 %	15,98 %	29,08 %
HGH		0,55 %	13,42 %	27,36 %	54,35 %
HGH	X	0,53 %	6,38 %	12,46 %	27,15 %

Tabelle 4: Wortfehlerraten für „Aurora-5“ bei Aufnahme in einer gestörten räumlichen Umgebung

Man stellt fest, dass die Fehlerraten bei Verwendung der Adaption der Referenzmuster in allen Fällen geringer sind im Vergleich zur Verwendung robuster akustischer Merkmale. Dies ist darauf zurückzuführen, dass bei dem Adaptionsverfahren auch eine Anpassung der Referenzmuster an die hallige Umgebung eines Raumes stattfindet. Eine Kompensation des Einflusses von Nachhall auf das Sprachsignal ist bei dem von ETSI standardisierten Verfahren nicht vorgesehen.

Zur Erkennung deutscher Ziffernketten werden die gleichen Aufnahmen der „RVG“ Sprachdatensammlung verwendet, die auch für die Experimente zur Erkennung im Auto benutzt werden. Die Aufnahme im Freisprechmodus wird durch die Faltung jedes Sprachsignals mit einer in einem Besprechungsraum bestimmten Impulsantwort realisiert. Die Impulsantwort wurde aktuell im Rahmen dieses Projekts erzeugt. Die Vorgehensweise zur Bestimmung der Impulsantwort wird im Anhang erläutert. Als Störgeräusch wurden die gleichen, typischerweise in

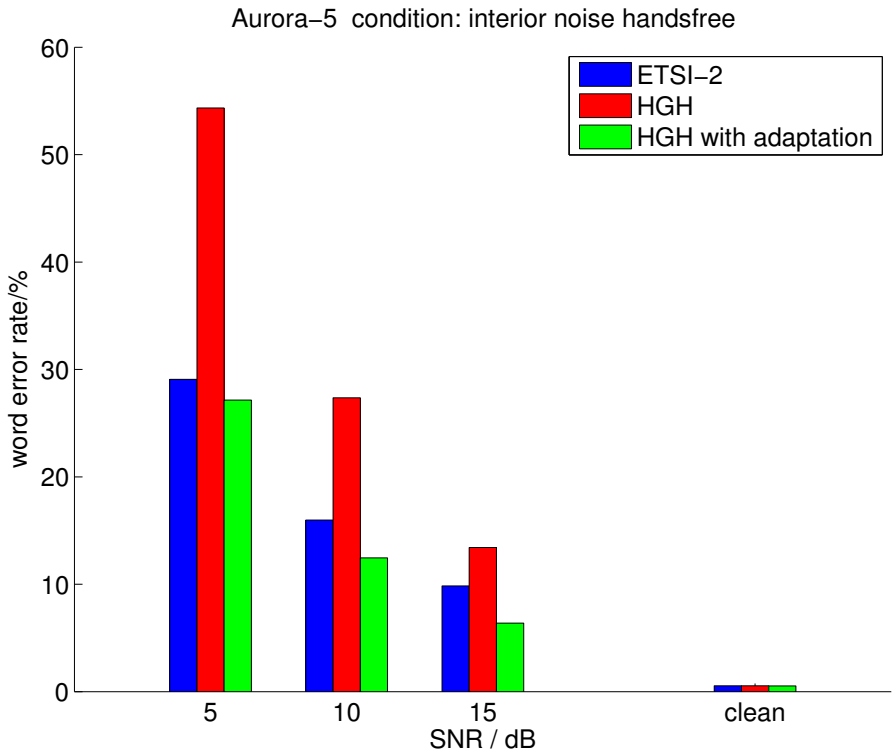

Bild 4: Wortfehlerraten für „Aurora-5“ bei Aufnahme in einer gestörten räumlichen Umgebung

einer räumlichen Umgebung auftretenden Hintergrundstörungen additiv überlagert, die auch bei den „Aurora-5“ Experimenten benutzt werden.

Als Referenzmuster werden die gleichen, aus ungestörten Sprachdaten trainierten, geschlechtsspezifischen HMMs wie bei den Experimenten zur Erkennung im Auto verwendet. Tabelle 5 kann man die Wortfehlerraten für die verschiedenen Analyseverfahren sowie für die Erkennung mit Adaption entnehmen. Jede Wortfehlerrate resultiert aus der Wortketten-Erkennung von 4767 Äußerungen mit 19947 gesprochenen Ziffern. Die gleichen Ergebnisse werden in Bild 5 graphisch veranschaulicht.

		SNR/dB			
Analyseverfahren	Adaption	Clean	15	10	5
ETSI-2		3,18 %	33,56 %	40,71 %	52,99 %
HGH		3,33 %	37,55 %	48,76 %	64,23 %
HGH	X	3,43 %	26,99 %	37,15 %	53,91 %

Tabelle 5: Wortfehlerraten für „RVG“ bei Aufnahme in einer gestörten räumlichen Umgebung

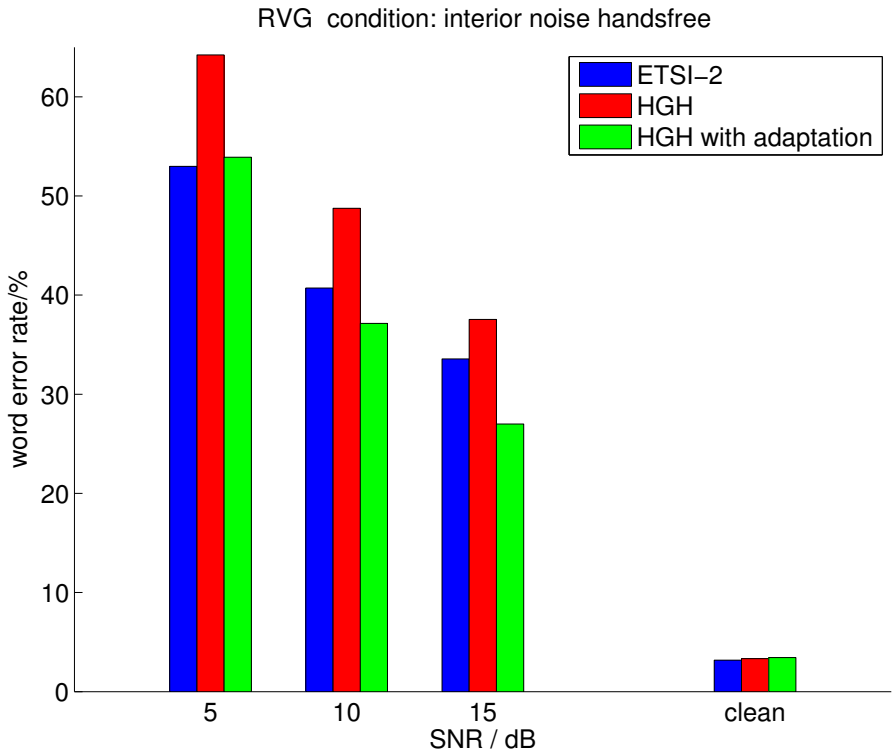

Bild 5: Wortfehlerraten für „RVG“ bei Aufnahme in einer gestörten räumlichen Umgebung

Bei einem Vergleich der Wortfehlerraten mit den bei der Erkennung englischer Ziffernketten erzielten Ergebnissen fällt auf, dass die Fehlerraten bei der Erkennung der deutschen Ziffernketten deutlich höher sind. Dafür sind im Wesentlichen 2 Gründe anzuführen. Wie schon zuvor erwähnt, weisen die ungestörten RVG Daten ein schlechteres SNR auf. Daraus resultieren prinzipiell höhere Fehlerraten. Zudem wurde bei der Erzeugung der Aurora Daten das Freisprechen in einem Raum mit einer mittleren Nachhallzeit von etwa 0,45 s simuliert, wohingegen die zur Simulation des Freisprechens bei den RVG Daten verwendete Raumimpulsantwort in einem Raum mit einer Nachhallzeit von etwa 0,7 s bestimmt wurde. Die größere Nachhallzeit führt zu einer deutlichen Erhöhung der Fehlerraten.

Neben der Erkennung der künstlich gestörten Sprachsignale der Aurora-5 und der RVG Datenbasen wird die Erkennung von real in Räumen aufgenommenen Signalen untersucht. Dazu werden die italienischen Daten der Sprachdatensammlung mit der Bezeichnung „SpeeCon“ verwendet, die von dem industriellen Partner im Rahmen dieses Projekts zur Verfügung gestellt wurden. Es wird die Erkennung von 19 jeweils isoliert gesprochenen Wörtern betrachtet. Die Beschränkung auf die ausgewählten Wörter ergab sich aus der Forderung, eine genügende Anzahl

von Äußerungen zum Training von Referenzmodellen zur Verfügung zu haben. Eine Zusammenstellung der Wörter findet sich in Tabelle 6.

arresta	calcolatrice	cartella	casse	centro
citta	cultura	lavoro	macchina	mostra
no	numero	opzioni	piu	ristorante
satellite	seleziona	si	televisione	

Tabelle 6: Verwendete „SpeeCon“ Wörter

Die Aufnahmen wurden gleichzeitig mit 4 Mikrofonen in einer räumlichen Umgebung erstellt. Die Sprecher benutzten ein Headset zur Nachbesprechung. Daneben wurden die Signale von 3 Mikrofonen aufgezeichnet, die an verschiedenen Stellen mit einem zunehmenden Abstand zum Sprecher positioniert wurden. Die im Nahbesprechungsmodus aufgezeichneten Sprachsignale werden zum Training von 2 geschlechtsspezifischen HMMs für jedes zu erkennende Wort verwendet. Jedes HMM besitzt 16 Zustände. Das Auftreten jedes akustischen Parameters in einem Zustand wird mit einer Mischung zweier Gauß-Verteilungen modelliert. Zur Erkennung der isoliert gesprochenen Wörter werden die Aufnahmen mit den 4 Mikrofonen separat je Mikrofon betrachtet. Dabei stellen sich die in Tabelle 7 zusammengestellten Wortfehlerraten ein.

Analyseverfahren	Adaption	Nahbesprechung	Position-1	Position-2	Position-3
ETSI-2		0,51 %	5,50 %	24,92 %	48,65 %
HGH		1,85 %	17,23 %	53,14 %	60,61 %
HGH	X	1,07 %	4,49 %	22,22 %	39,17 %

Tabelle 7: Wortfehlerraten für die Erkennung der isoliert gesprochenen „SpeeCon“ Wörter

Für einen zunehmenden Abstand des Mikrofons zum Sprecher von Position 1 zu Position 3 stellt sich eine deutliche Erhöhung der Fehlerraten ein. Bei einem Vergleich der beiden Vorgehensweisen zur Erhöhung der Robustheit stellt man fest, dass man mit der Adaption der Referenzmodelle deutliche niedrigere Fehlerraten erzielt im Vergleich zur Extraktion robuster Merkmale. Dies liegt, wie schon zuvor erwähnt, darin begründet, dass die Kompensation des Einflusses von Nachhall auf das Sprachsignal bei dem von ETSI standardisierten Verfahren nicht vorgesehen ist.

Anhang - Erkennungsergebnisse ohne Berücksichtigung der Aufnahme im Freisprechmodus

Aus dem ersten Projektabschnitt ergab sich die Fokussierung auf die Erkennung in Kraftfahrzeugen und in Räumen im Freisprechmodus. Dazu wurden in den vorhergehenden Abschnitten die entsprechenden Erkennungsergebnisse aufgeführt. Zu Vergleichszwecken sind allerdings auch die Ergebnisse interessant, die ohne Berücksichtigung einer Aufnahme im Freisprechmodus erzielt werden. Beispielsweise kann man durch den Vergleich der Erkennung von Äußerungen, die in der gleichen Störumgebung einmal ohne und einmal mit Berücksichtigung der Aufnahme im Freisprechmodus aufgezeichnet wurden, eine Aussage darüber machen, wie gut ein Verarbeitungsschritt zur Kompensation der Einflüsse des Freisprechens geeignet ist. Den nachstehenden Tabellen 8 bis 11 können dazu die Ergebnisse ohne Berücksichtigung des Freisprechens entnommen werden, die unmittelbar mit den in den Tabellen 1 und 2 sowie in den Tabellen 4 und 5 aufgeführten Wortfehlerraten verglichen werden können. Einer Äußerung der TIDigits oder der RVG Sprachdatensammlung wurde dabei jeweils der gleiche Abschnitt eines Störsignals für das vorgegebene SNR überlagert. Der einzige Unterschied ist die Simulation der Aufnahme im Freisprechmodus, die bei den nachstehenden Ergebnissen nicht berücksichtigt wurde.

		SNR/dB				
Analyseverfahren	Adaption	Clean	15	10	5	0
ETSI-2		0,55 %	1,35 %	2,52 %	5,94 %	16,14 %
HGH		0,55 %	5,42 %	16,12 %	39,52 %	70,99 %
HGH	X	0,53 %	1,15 %	2,11 %	5,83 %	18,73 %

Tabelle 8: „Aurora-5“ Wortfehlerraten für die gestörte Umgebung im Auto (ohne Freisprechen)

		SNR/dB				
Analyseverfahren	Adaption	Clean	15	10	5	0
ETSI-2		3,18 %	4,61 %	7,22 %	14,13 %	30,41 %
HGH		3,33 %	9,21 %	17,93 %	35,27 %	63,31 %
HGH	X	3,43 %	4,97 %	7,69 %	15,14 %	32,66 %

Tabelle 9: „RVG“ Wortfehlerraten für die gestörte Umgebung im Auto (ohne Freisprechen)

		SNR/dB			
Analyseverfahren	Adaption	Clean	15	10	5
ETSI-2		0,55 %	2,48 %	5,38 %	13,10 %
HGH		0,55 %	5,70 %	16,66 %	43,15 %
HGH	X	0,53 %	2,14 %	4,89 %	13,81 %

Tabelle 10: „Aurora-5“ Wortfehlerraten für eine gestörte räumliche Umgebung (ohne Freisprechen)

		SNR/dB			
Analyseverfahren	Adaption	Clean	15	10	5
ETSI-2		3,18 %	7,54 %	13,45 %	25,19 %
HGH		3,33 %	9,68 %	18,81 %	37,97 %
HGH	X	3,43 %	8,60 %	15,31 %	30,20 %

Tabelle 11: „RVG“ Wortfehlerraten für eine gestörte räumliche Umgebung (ohne Freisprechen)

Anhang – Messtechnische Bestimmung von Raumimpulsantworten

Von dem industriellen Partner dieses Projekts, der Firma Teleca, wurden die italienischen Aufnahmen, die im Rahmen des durch die EU geförderten Projekts „SpeeCon“ erstellt wurden, zur Verfügung gestellt. Das Ziel des SpeeCon Projekts war eine Aufnahme von Sprachdaten in verschiedenen akustischen Umgebungen, um sie zum Training von Spracherkennungssystemen und zur Durchführung von Erkennungsexperimenten verwenden zu können. Dabei wurden die Äußerungen einer Person gleichzeitig mit 4 Mikrofonen aufgenommen. Da die Firma Teleca bzw. eine frühere Arbeitsgruppe der Firma Ericsson, die komplett in die Ericsson beratende Firma Teleca übergegangen ist, aktiv an der Erstellung der SpeeCon Aufnahmen beteiligt war, sind dort zwei komplette Messaufbauten vorhanden, die für die Aufnahmen eingesetzt wurden. Über die ursprüngliche Planung des Projekts hinausgehend, hat der industrielle Partner einen Aufbau zur Durchführung von Messungen der Arbeitsgruppe an der Hochschule Niederrhein zur Verfügung gestellt. Zu einer Einführung in dieses Aufnahme- und Messsystem und zur Übernahme besuchte der wissenschaftliche Mitarbeiter, Herr A. Kitzig, im Zeitraum vom 19.11.08 bis zum 21.11.08 die Firma Teleca.

Das Aufnahmesystem besteht aus einer transportablen und netzunabhängig zu betreibenden Anordnung von Vorverstärkern, an die bis zu 4 Mikrofone angeschlossen werden können. Die verstärkten Mikrofonsignale können dann mit Hilfe hochwertiger A/D-Wandler direkt in digitaler Form auf einem Laptop aufgezeichnet werden.

Neben der Verwendung zur Aufnahme von Sprachsignalen war in einem Abschnitt des SpeeCon Projekts auch die Erstellung eines Messaufbaus und einer entsprechenden Software zur Bestimmung der Raumimpulsantwort vorgesehen, um damit die raumakustischen Bedingungen bei Aufnahmen im Freisprechmodus zu bestimmen und diese in späteren Simulationsexperimenten verwenden zu können. Der für die Erstellung der in Matlab geschriebenen Software zuständige SpeeCon Projektpartner hat der Arbeitsgruppe an der Hochschule Niederrhein diese Software freundlicherweise zur Verfügung gestellt und deren Benutzung zur eigenen Bestimmung von Impulsantworten gestattet.

Der Messaufbau ist in Bild 6 dargestellt. Mit Hilfe eines CD-Players werden über einen Lautsprecher zeitlich versetzt ein Rosa-Rauschen und eine Maximum-Length-Sequenz (MLS) wiedergegeben. Ein MLS Signal ist ein pseudozufälliges Rauschsignal, das aufgrund seiner Korrelationseigenschaften sehr gut zur Bestimmung von Impulsantworten geeignet ist. Der

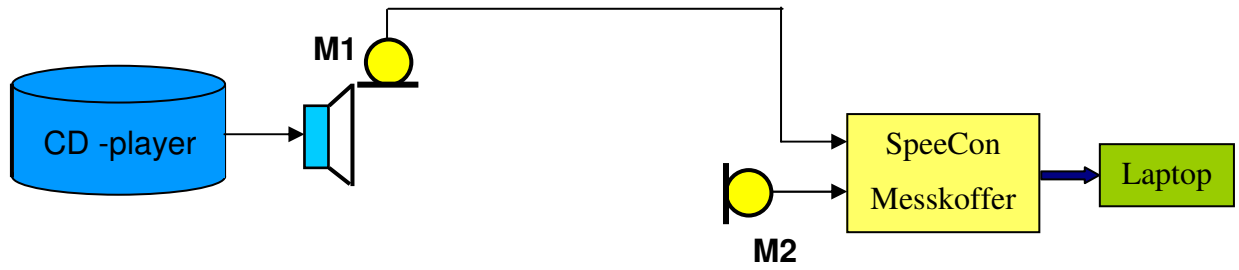

Bild 6: Messaufbau zur Bestimmung von Raumimpulsantworten

Lautsprecher wird an der Stelle im Raum positioniert, an der sich der an der Sprachdatensammlung teilnehmende Sprecher befindet. Mit Hilfe von Mikrofon M1 wird das vom Lautsprecher abgestrahlte Signal direkt aufgezeichnet. Das zweite Mikrofon M2 wird an der Stelle im Raum positioniert, an der sich auch das bei der Sprachdatensammlung verwendete Aufnahmemikrofon befindet. Die beiden Mikrofonsignale werden synchron mit Hilfe des SpeeCon Messaufbaus auf einem Laptop gespeichert. Im Anschluss kann mit Hilfe der entwickelten Software aus den beiden Mikrofonsignalen die Raumimpulsantwort geschätzt werden.

Der Messaufbau wurde im Rahmen dieses Projekts dazu benutzt, um in einem Kraftfahrzeug des Typs VW Touran und in einem Besprechungsraum verschiedene Impulsantworten zu bestimmen. Zwei der dabei ermittelten Impulsantworten werden in Bild 7 und in Bild 8 dargestellt.

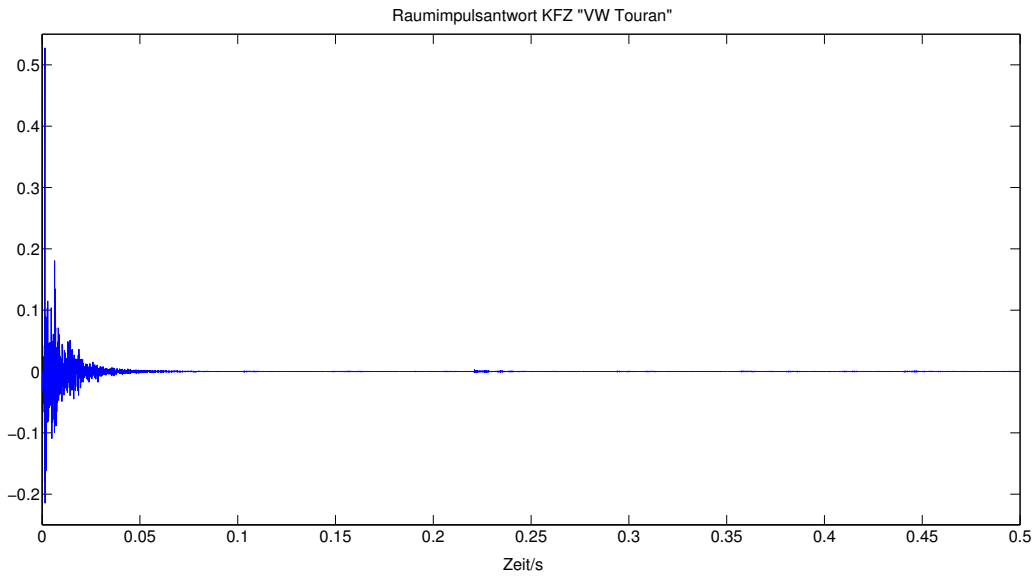

Bild 7: Raumimpulsantwort in einem Kraftfahrzeug des Typs VW Touran

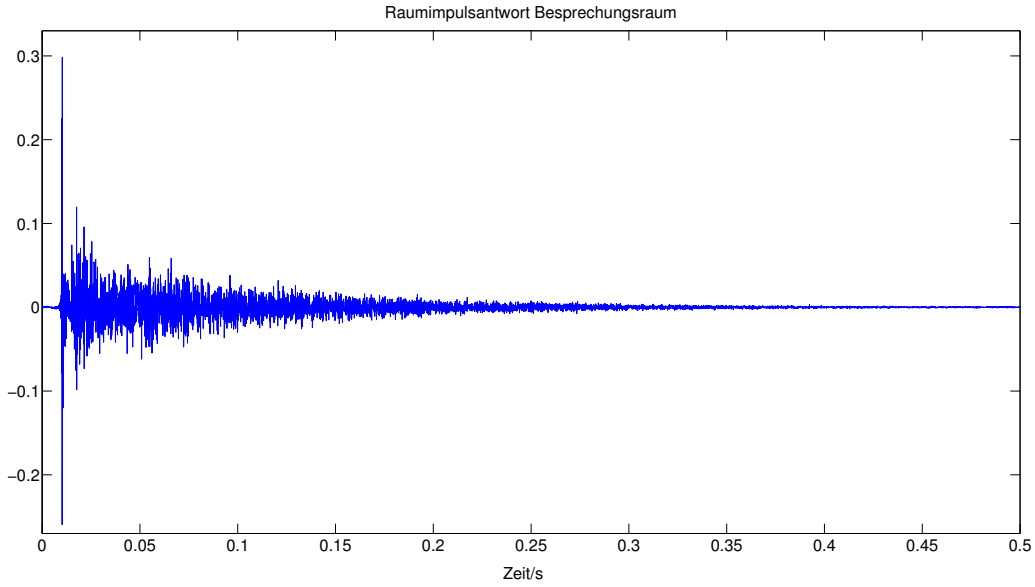

Bild 8: Raumimpulsantwort in einem Besprechungsraum (T_{60} ~ 0,7 s) bei einem Abstand von ca. 3,5 m zwischen Lautsprecher und Mikrofon

Zur Bestimmung der Impulsantwort in dem Auto wurde der Lautsprecher auf dem Fahrersitz in Kopfhöhe positioniert, um die Situation eines freisprechenden Fahrers nachzuempfinden. Das Mikrofon wurde in der Nähe des Rückspiegels oberhalb der Frontscheibe angebracht.

Die Impulsantwort im Innern des Kraftfahrzeugs ist wesentlich kürzer als die im Besprechungsraum, da das Raumvolumen des Autos wesentlich geringer ist als das des Raumes. Der erste größere Impuls tritt bei der Impulsantwort des Raumes etwa nach einer Zeit von 10 ms auf, was der Laufzeit des Schalls auf dem direkten Weg vom Lautsprecher zum Mikrofon entspricht. Die beiden dargestellten Impulsantworten wurden dazu verwendet, um den Freisprechmodus bei den RVG Sprachdaten zu simulieren. Die „ungestörten“ Sprachsignale werden dazu mit den Impulsantworten gefaltet, bevor die Störung gemäß dem gewünschten SNR hinzuaddiert wird.

Anhang – Aufnahme von Sprachdialogen bei Freisprechen im Auto

Da zur Bestimmung der Raumimpulsantwort der im vorherigen Abschnitt erwähnte Messaufbau ohnehin in ein Kraftfahrzeug eingebracht wurde, wurde von Seiten des industriellen Partners, der Firma Teleca, der Wunsch geäußert, die Aufnahmen einiger „kontrolliert“ verlaufender Sprachdialoge in dem stehenden und in dem fahrenden Kraftfahrzeug aufzuzeichnen. Solche Aufnahmen werden dazu benötigt, um die bei der Firma Teleca entwickelten

Freisprecheinrichtungen und die dabei eingesetzten Verfahren zur Echokompensation zu testen und zu verbessern. Der kontrollierte Ablauf bezieht sich dabei auf eine zeitliche Definition der Sprachpausen zwischen den Äußerungen des „nahen“ und des „fernen“ Sprechers sowie im Sonderfall des gleichzeitigen Sprechens von nahem und fernem Sprecher, der sogenannten „Double-talk“ Situation, der Festlegung der zeitlichen Dauer dieser Phase. Durch eine einfache Modifikation des bei der Arbeitsgruppe an der Hochschule Niederrhein ohnehin vorhandenen Sprachdialogsystems konnte dem Wunsch Rechnung getragen werden.
Es wurde der in Bild 9 dargestellte Messaufbau verwendet.

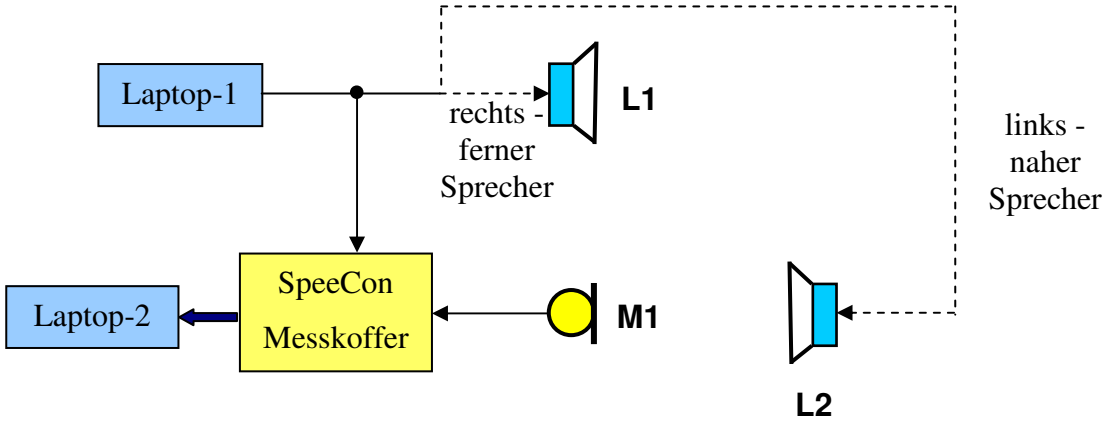

Bild 9: Aufbau zur Generierung und Aufzeichnung „kontrolliert“ verlaufender Sprachdialog

Der Laptop-1 wird dazu verwendet, den kompletten Dialog zwischen einem nahen und einem fernen Sprecher wiederzugeben. Der nahe Sprecher wird über einen der beiden Stereokanäle (z.B. den linken Kanal), der ferne Sprecher über den anderen (z.B. rechten) Kanal wiedergegeben.
Zur Erzeugung dieses Dialogs wird eine Dialogcontroller Software eingesetzt, die eigentlich zur Realisierung eines Dialogs mit einem auf einer Spracherkennung basierenden automatischen System verwendet wird. Es wird in diesem Fall „nur“ die Sprachausgabe des Dialogsystems benutzt, wobei diese um die Wiedergabe von Stereosignalen bei einer Abtastfrequenz von 44,1 kHz erweitert und modifiziert wurde. Die Software benötigt die Beschreibung des Dialogs als eine Folge von Zuständen. Ein Zustand definiert die akustische Wiedergabe eines Audiofiles, das in der Regel entweder eine Sprachausgabe des nahen oder des fernen Sprechers beinhaltet. Eine Ausnahme stellt die „double-talk“ Situation dar, bei der ein Audiofile die sich überlappenden Sprachausgaben des nahen und des fernen Sprechers beinhaltet. Zwischen den Sprachausgabezuständen können „Pause“-Zustände eingefügt werden, mit denen die zeitliche

Dauer zwischen dem Ende einer Äußerung und dem Beginn der nächsten Äußerung definiert werden kann.

Als sprachliche Ausgaben von nahem und fernem Sprecher wurden von Hörbuch CDs einzelne oder mehrere hintereinander gesprochene Sätze extrahiert, so dass sprachliche Äußerungen mit einer Dauer von ca. 3 bis 11 Sekunden vorhanden sind. Es wurden jeweils 12 Ausschnitte von 3 verschiedenen männlichen und 3 weiblichen Sprechern extrahiert. Dabei wurde darauf geachtet, dass keine längeren Sprachpausen am Beginn und am Ende jeder Äußerung vorhanden sind, um die Äußerungen zur Erzeugung von „double-talk“ Abschnitten mit einer definierten Dauer des „double-talk“ verwenden zu können. Die extrahierten Abschnitte wurden als Monosignale und nach einer Skalierung des Signals auf den kompletten 16 Bit-Wertebereich mit der Abtastfrequenz von 44,1 kHz im WAV Format abgespeichert. Zur Skalierung wird der betragsmäßig größte Amplitudenwert bestimmt. Die Skalierung wird dann als Multiplikation mit einem Verstärkungsfaktor vorgenommen, so dass der betragsmäßig größte Amplitudenwert dem kleinsten oder größten möglichen Wert des 16 Bit Bereichs entspricht. Das Extrahieren, Umwandeln in ein Monosignal und das Skalieren wurde mit der Software „Adobe Audition“ durchgeführt.

In Matlab wurde eine Software erstellt, um die Monosignale in Stereosignale umzuwandeln, bei denen das Signal nur in einem Kanal vorhanden ist. Der andere Kanal beinhaltet Nullwerte. Des Weiteren wurde die Möglichkeit geschaffen, um aus 2 Signalen ein „double-talk“ Signal mit einer definierten Dauer des „double-talk“ zu erzeugen. Mit Hilfe der Dialogdefinition und der Matlab Funktionen können gewünschte Dialoge oder Dialogsituationen auf einfache Weise definiert und modifiziert werden.

Zur Wiedergabe des nahen Sprechers wird ein hochwertiger Lautsprecher (der Firma Fostex) eingesetzt, der sich auf dem Beifahrersitz etwa in Höhe des Mundes eines sitzenden Sprechers befindet. Der Lautstärkepegel wird in etwa so eingestellt, dass er dem Pegel eines realen Sprechers während der geräuscherfüllten Fahrsituation entspricht.

Zur Wiedergabe des fernen Sprechers wurden alternativ verschiedene Lautsprecher eingesetzt. Dabei handelt es sich um einen tatsächlich bei einer Freisprecheinrichtung mitgelieferten Lautsprecher als auch geringfügig hochwertigere Lautsprecher. Der jeweilige Lautsprecher wird an den Stellen positioniert, die bei Kraftfahrzeugen dafür typischerweise gewählt werden. Der

Pegel des fernen Sprechers wird so eingestellt, dass der Fahrer des Fahrzeugs den fernen Sprecher in der geräuscherfüllten Fahrsituation gut verstehen kann.

Zur Aufnahme wird der „SPEECON“ Messkoffer in Kombination mit Laptop-2 verwendet. Auf einem Kanal wird das elektrische Lautsprechersignal des fernen Sprechers aufgezeichnet. Auf dem zweiten Kanal wird synchron das über ein Mikrofon aufgenommene Signal aufgezeichnet. Das Mikrofon wurde im Bereich des Rückspiegels oberhalb der Frontscheibe positioniert. Es wurden alternativ verschiedene Mikrofone verwendet, von typischerweise bei einer Freisprecheinrichtung eingesetzten Mikrofontypen bis hin zu hochwertigeren Mikrofonen.

Um die Aufnahmen während der Fahrt durchzuführen, muss neben dem Fahrer und dem hochwertigen Lautsprecher als „Beifahrer“ eine weitere Person im Fond des Fahrzeugs sein, die den Wiedergabe-Laptop und den Aufnahme-Laptop bedient. Es wurden erfolgreich mehrere Sprachdialoge in einem VW Touran aufgezeichnet. Dabei erfolgten die Aufnahmen in verschiedenen Fahrsituationen im Stadtverkehr und auf der Autobahn.