Ambiente Sprachsteuerung für einen Persönlichen Aktivitäts- und Haushaltsassistenten

Ambient Voice Control for a Personal Activity and Household Assistant

Niko Moritz, Stefan Goetze und Jens-E. Appell
Fraunhofer Institut für Digitale Medien Technologie (IDMT), Projektgruppe Hör-, Sprach- und Audiotechnologie (HSA), Marie-Curie-Str. 2, 26129 Oldenburg, Deutschland. E-Mail: niko.moritz@idmt.fraunhofer.de

Kurzfassung

Mit der Hilfe von assistiven Technologien kann die Lebensqualität älterer oder beeinträchtigter Menschen verbessert werden. In diesem Beitrag wird die Nutzung von automatischen Spracherkennungssystemen (automatic speech recognition, ASR) als natürliche Schnittstelle zur Steuerung von Technologien diskutiert, die vornehmlich ältere Nutzer im Alltag unterstützenden können. Hierbei wird insbesondere auf die Nutzung von Freisprecheinrichtungen, den damit verbundenen Herausforderungen für die Spracherkennungssoftware und den erzielten Nutzen für ältere Menschen eingegangen. Aktuelle Ansätze zur Verbesserung der Robustheit werden vorgestellt, diskutiert und mit einem ASR Experiment veranschaulicht.

Abstract

Technologies for ambient assisted living (AAL) are used to increase the quality of life of older or impaired persons. This contribution discusses the utilization of automatic speech recognition (ASR) as a natural interface for controlling assistive technologies in everyday life situations. We focus on the use of hands-free systems, the technical challenges for the ASR software caused by this and the benefits for older persons. Moreover, state-of-the-art approaches for improving robustness of ASR systems are presented, discussed and demonstrated by an ASR experiment.

1 Einleitung

Der demografische Wandel, der zu einem wachsenden Anteil älterer Menschen in der Bevölkerung führen wird [1]-[3], macht neue technische Lösungen erforderlich, um den Anforderungen z. B. im Bereich der sozialen Betreuung und Pflege gerecht zu werden. Assistive Technologien zeigen großes Potenzial, um ältere Menschen im Alltag zu unterstützen und ihnen damit ein längeres unabhängiges Leben zu ermöglichen [4]. In diesem Zusammenhang haben Informations- und Kommunikationstechnologien große soziale Relevanz. Doch besonders für ältere Menschen ist diese Art der Technologie aufgrund der oft hohen Komplexität nicht einfach zu bedienen. Daher ist der Entwurf einfacher und intuitiver Nutzerschnittstellen, insbesondere für technische Systeme die sich an ältere Menschen richten, von hoher Relevanz [4]-[6].

Die Verwendung von Sprache ist die natürlichste Form der menschlichen Kommunikation. Daher ist auch die Steuerung assistiver Technologien durch Sprache wünschenswert und bietet die Möglichkeit der intuitiven und nutzerfreundlichen Bedienung von assistiven Systemen, zum Beispiel in der häuslichen Umgebung. Darüber hinaus bietet die Spracheingabe insbesondere in Situationen in denen die Hände nicht frei verwendet werden können große Vorteile gegenüber konventionellen Eingabemethoden wie bspw. einer Fernbedienung oder einer Maus und Tastatur [4][7].

Die robuste automatische Spracherkennung (automatic speech recognition, ASR) in einem Freisprechszenario stellt in schwierigen akustischen Umgebungen jedoch noch ein bisher nicht zufriedenstellend gelöstes Problem dar [8][9]. Die Zuverlässigkeit (bzw. Robustheit) aktueller Spracherkennersysteme ist noch immer weit von der Leistung des Menschen entfernt [10]. Aus diesem Grund werden in diesem Beitrag Ergebnisse einer Nutzerstudie präsentiert, in der u.a. Toleranzgrenzen älterer Nutzer im Bezug auf fehlerhafte Spracherkennung ermittelt wurden. Außerdem wurde die allgemeine Akzeptanz eines solchen Systems durch die Nutzer überprüft (vgl. Abschnitt 2).

Zur Verbesserung der Erkennungsleistung eines ASR Systems in verrauschter und/oder halliger Umgebung existieren verschiedene Ansätze. Zum einen können Konzepte eingesetzt werden, die das Signal-zu-Rauschverhältnis (signal-to-noise ratio, SNR) erhöhen (vgl. Abschnitt 3.1 und die Referenzen darin). Diese Signalverarbeitungskonzepte können zur Vorverarbeitung der Spracheingabe von ASR Systemen genutzt werden, um so die Robustheit zu verbessern, insbesondere wenn Freisprecheinrichtungen verwendet werden. Es sollte jedoch bedacht werden, dass diese Konzepte unter Umständen zu spektralen Verzerrungen führen, welche die akustischen Merkmale auf denen

ISBN 978-3-8007-3323-1

ein Spracherkenner basiert derart stören, dass dies zu einer Verschlechterung der Erkennerleitung führen kann, obwohl sich das SNR verbessert hat [9]. Daher muss der Nutzen verschiedener Vorverarbeitungskonzepte immer im Zusammenhang mit dem ASR System betrachtet werden.
Neben einer geeigneten Vorverarbeitung existieren zum anderen Konzepte, welche die Erkennungsleistung eines ASR Systems in gestörten Umgebungen selbst verbessern können. Aktuelle Methoden hierzu werden in Abschnitt 3.2 vorgestellt und diskutiert. Eine einfache Methode ist es bspw. einen strukturierten Dialog zu verwenden, anstatt von kontinuierlicher Sprache. Dies hat zwei Vorteile. Einerseits kann dadurch die Komplexität der Spracherkennungssoftware gering gehalten werden und zum anderen wird die Steuerung eines solchen Systems durch einzelne gesprochene Kommandos von vielen Nutzern bevorzugt [11]. Als Hauptgrund für einen strukturierten Dialog wird in [11] z. B. angeführt, dass diese Art der Eingabe unmissverständlich und bereits von anderen Anwendungen bekannt ist.
In Abschnitt 4 dieses Beitrags werden Ergebnisse eines ASR Experiments vorgestellt, welches in einem AAL Laborraum durchgeführt wurde. In diesem Raum wurde ein Deckenmikrofon in einiger Entfernung zum Sprecher und ein direkt am Mund des Sprechers positioniertes („close-talk“) Mikrofon benutzt. Die Erkennerleistung des ASR Systems wurde im Bezug auf das verwendete Mikrofon für die Trainings- und Testaufnahmen evaluiert.

2 Studie zur Nutzerakzeptanz

In [11] wird eine Nutzerstudie beschrieben die durchgeführt wurde, um die Toleranzgrenze und Akzeptanz der Nutzer für fehlerhafte Kommandoerkennungen bei Spracheingabesystemen zu evaluieren. Für diesen Test wurden insgesamt 12 Probanden im Alter zwischen 63 und 75 Jahren befragt. Das Experiment wurde unter der Verwendung eines Mock-Up Systems durchgeführt, welches eine Kalenderanwendung als Teil eines persönlichen Aktivitäts- und Haushaltsassistenten (PAHA) simulierte [12]. Das bedeutet, dass die Ein- und Ausgaben des Systems von einem Versuchsleiter kontrolliert wurden, der auf die Kommandos der Versuchspersonen regierte. Der Versuchsleiter hat somit die Aufgaben des ASR Systems übernommen, wodurch vorher definierte Erkennungsfehler gezielt eingebracht werden konnten. Die Systemausgabe wurde akustisch durch vorher aufgenommene Sätze präsentiert. Mit diesem Vorgehen wurden drei vordefinierte Testphasen durchlaufen, um die Toleranzgrenze der Nutzer im Bezug auf Falscherkennungen zu evaluieren. In jeder dieser Testphasen konnte der Nutzer einen neuen Termin durch Sprachkommandos in das Kalendersystem eingeben. Im Anschluss an den Test wurden die Probanden nach ihrem Eindruck zum Kalendersystem befragt.
In der ersten Testphase lief das ASR System fehlerfrei. Dadurch sollten sich die Probanden an das System gewöhnen und einen Eindruck davon bekommen, wie das System im prinzipiell funktioniert. Anschließend wurde in einer zweiten Testphase ein Erkennungsfehler eingebaut, der dazu führte, dass die Testperson das falsch erkannte Kommando durch Wiederholen der Spracheingabe korrigieren musste. In der dritten und letzten Testphase wurde eine Schleife von Endlosfehlern durch den Versuchleiter simuliert, wodurch die Frustrationsgrenze des Probanden ermittelt werden sollte. Um diese festzustellen wurden eine gelbe und eine rote Karte an die Probanden ausgeteilt. Mit der gelben Karte konnte eine erste Verwarnung für das System angezeigt werden, die als erstes Anzeichen der Frustration gewertet wurde. Das Ziehen der roten Karte führte zum Abbruch des Tests. Mit diesem Vorgehen wurde die Toleranzgrenze der Probanden für Fehler durch das ASR System bestimmt.
Nach jeder Phase wurden die Eigenschaften „Nutzerfreundlichkeit“, „Intuitivität der Bedienung“, „Verständlichkeit“, „Nützlichkeit“ und „Akzeptanz“ ausgewertet. Diese Eigenschaften wurden auf einer Skala von 1 („überhaupt nicht zutreffend“) bis 5 („voll zutreffend“) bewertet. Das Ergebnis nach der ersten Phase war, dass die Nutzerfreundlichkeit und Akzeptanz als sehr gut bewertet wurden (zwischen 4 und 5 von allen 12 Testpersonen). Das Ergebnis im Bezug auf Nützlichkeit wurde ebenfalls als gut bewertet (zwischen 3 und 5). Die Bewertung der Eigenschaften Verständlichkeit und Intuitivität waren dagegen etwas durchwachsener. Der interessierte Leser wird für eine ausführliche Diskussion der Ergebnisse auf Referenz [11] verwiesen.
Nach der zweiten Testphase (mit einem eingebauten Fehler) hat sich die Bewertung des Systems durch die Probanden nicht signifikant verändert.
Die Evaluation der Frustrationsgrenze in der dritten Testphase führte zu interessanten Ergebnissen. Die Anzahl der Wiederholungen der falsch erkannten Kommandos bis zum Abbruch des Tests durch den Probanden selbst reichten von 0 bis 17. Der Medianwert hierfür ist 6. Wie in [11] angegeben wird, wurde eine solch hohe Toleranzgrenze von den Versuchsleitern für dieses Experiment nicht erwartet. Neun der zwölf Probanden gaben an, dass sie ein solches System in der Zukunft nutzen würden, auch wenn die Bereitschaft das System zum Zeitpunkt der Befragung zu nutzen eher gering war. Die genannten Gründe waren, dass das präsentierte Mock-Up System noch sehr begrenzt in den Eigenschaften und den Ausgabemodalitäten war, da z. B. nur eine akustische Ausgabe der Systemantwort zur Verfügung stand. Die Referenzen [7] und [11] bieten eine detaillierte Übersicht über muli-modale Ausgabemöglichkeiten eines persönlichen Aktivitäts- und Haushaltsassistenten.
Außerdem gaben viele Nutzer an das System mit einem Schlüsselwort oder durch Klatschen aktivieren zu wollen. Die Referenz [13] beschreibt für diesen Zweck Methoden der akustischen Ereigniserkennung für Anwendungen im Bereich Ambient Assisted Living.

ISBN 978-3-8007-3323-1

3 Heraus- und Anforderungen an ambiente sprachgesteuerte assistive Technologien

Die robuste automatische Spracherkennung zusammen mit entfernten Raummikrofonen, d.h. nicht für den sogenannten „close-talk“ Fall, ist noch immer eine große Herausforderung und aktuelles Forschungsthema im Bereich der automatischen Spracherkennung. Der Grund hierfür ist die räumliche Distanz zwischen den verwendeten Mikrofonen und der Quelle des gewünschten Sprachsignals, welches im Allgemeinen der Mund eines Sprechers ist. Dadurch nehmen die Mikrofone das gewünschte Sprachsignal nicht direkt auf, sondern auch Umgebungsgeräusche und verzögerte Versionen des Ursprungssignals (Raumhall), welche stark von der Umgebung abhängen in der das System eingesetzt wird.

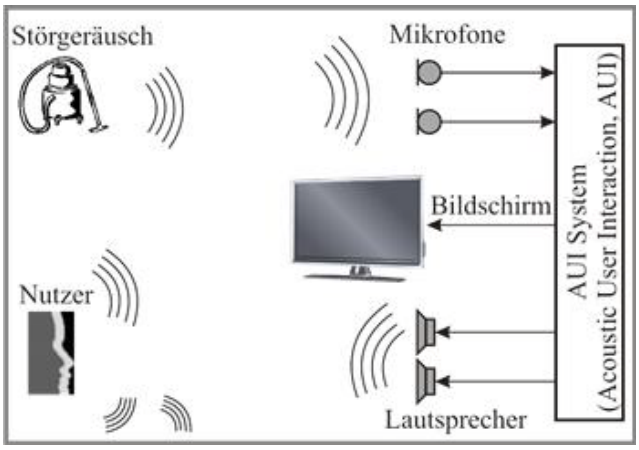

Abbildung 1: Schema der technischen Struktur und der Umgebungsbedingungen für einen persönlichen Aktivitäts- und Haushaltsassistenten. Das akustische Nutzer Interaktionssystem (acoustic user interaction framework, AUI) umfasst das ASR System und die Intelligenz, um die Ein- und Ausgabe in Abhängigkeit der spezifischen Anwendung zu steuern.

Abbildung 1 stellt ein Nutzerszenario dar, wie es zu Hause als ein ambientes assistives Gerät installiert sein könnte. Das dargestellte Beispiel zeigt eine Situation in der die Mikrofone des Systems verschiedene Signale von verschiedenen Signalquellen empfangen. Auf der einen Seite ist dies das gewünschte Sprachsignal des Sprechers, der die Sprachbefehle darbietet, und auf der anderen Seite sind dies ungewünschte Störanteile (Rauschen), die das Sprachsignal stören. Die dargestellten Störquellen in Abbildung 1 sind ein Staubsauger und die Lautsprecher des System selbst, welche die akustische Ausgabe des Systems abspielen. Die Signalteile die von den Lautsprechern abgespielt und von den Mikrofonen wieder aufgenommen werden, sind allgemein als akustische Echos bekannt [14]. Die Kommandoeingabe eines Nutzers wird demzufolge stark gestört durch (i) Umgebungsgeräusche, (ii) akustische Echos und (iii) Raumhall, d.h. Echos der Spracheingabe die durch Reflexionen an Wänden verursacht werden, wie es z. B. von Sprache in einer Kirche bekannt ist.
In Abschnitt 3.1 werden verschiedene Signalverarbeitungskonzepte beschrieben, um das Nutzsignal zu verstärken und Störsignale zu unterdrücken. In Abschnitt 3.2 wird eine Standardmethode vorgestellt, welche eingesetzt wird, um die Erkennungsleistung eines ambienten ASR Systems in akustisch schwierigen Umgebungen verbessern zu können. Des Weiteren wird ein Einblick in ein aktuelles Forschungsfeld in der ASR dargeboten.

3.1 Vorverarbeitungskonzepte zur Verbesserung der Sprachqualität

Wie in Abbildung 1 dargestellt, können mehrere Mikrofone sowie Lautsprecher zur Schallaufnahme und zum Abspielen in einem Freisprechszenario für ASR verwendet werden. Die Mikrofonsignale werden von der Signalverarbeitungseinheit AUI (acoustic user interaction framework, vgl. Abbildung 1) verarbeitet, welche das gewünschte Signal verstärkt und mit Hilfe der ASR analysiert. Die Systemausgabe wird in diesem Beispiel akustisch über die Lautsprecher und/oder über einen Bildschirm dargeboten.

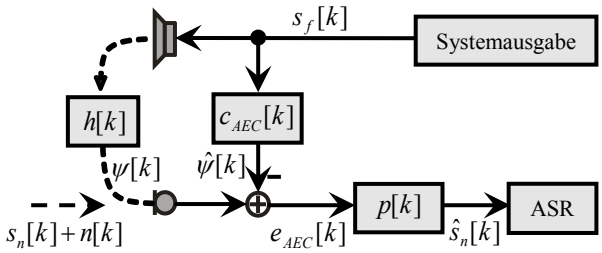

Abbildung 2: Signalverarbeitungsschema zur Unterdrückung ambienter Störgeräusche und akustischer Echos.

Abbildung 2 zeigt die Signalverarbeitungseinheit AUI für den einkanaligen Mikrofonfall im Detail. Das gewünschte Nutzsignal, d.h. die Spracheingabe des Nutzers, wird hier mit $s_n[k]$ gekennzeichnet. Ambiente Störgeräusche $n[k]$ und akustische Echos $\psi[k]$ sind Störungen für das ASR System, die mit dem Nutzsignal im Mikrofonkanal überlagert sind. Die akustische Ausgabe des Systems $s_f[k]$ wird über die Lautsprecher wiedergegeben. Aufgrund der akustischen Kopplung zwischen dem Lautsprecher und dem Mikrofon, werden Teile der akustischen Ausgabe wieder von dem Mikrofon aufgefangen. Zahlreiche Reflexionen des Signals an den Raumgrenzen (Wände, Boden und Decke) führen zu einer verhallten Version der Systemausgabe an der Position des Mikrofons. Mathematisch können diese Reflexionen durch die Raumimpulsantwort $h[k]$ beschrieben werden, wie in Abbildung 2 dargestellt. Da die Systemausgabe auch Sprachinformationen enthalten kann, würde dies ohne Unterdrückung zu einer starken Störung des ASR Systems führen. Sogenannte akustische Echokompensationsfilter $c_{\mathrm{AEC}}[k]$ schätzen den Signalanteil am Mikrofoneingang der aus dem Lautsprecher stammt, wodurch eine Unterdrückung des Selben möglich wird

ISBN 978-3-8007-3323-1

[15][16]. Stark hallige Umgebungen und mehrere Lautsprecher stellen eine besondere Herausforderung für die Echokompensation dar. Der interessierte Leser wird auf [14]-[18] für eine detaillierter Diskussion der technischen Anforderungen verwiesen.

Rest-Echos, die von der Echo-Kompensation nicht ausgelöscht werden konnten, sowie andere Störungen $n[k]$ die von dem Mikrofon aufgefangen wurden, sollen durch das nachfolgende Kompensationsfilter $p[k]$ unterdrückt werden, bevor das gefilterte Audiosignal anschließend mit einer ASR Software analysiert wird [9][19]. Kompensationsfilter unterdrücken allerdings generell nicht nur Störungen, sondern können eventuelle auch neue Störungen (sogenanntes „Musical Noise“) verursachen. Auch wenn diese Störungen klein in der Amplitude sein mögen, können sie die akustischen Merkmale auf denen ein ASR System basiert stark stören. Deshalb sollten Verzerrungen des gewünschten Signals so klein wie möglich gehalten werden [9].

Störungen aus verschiedenen räumlichen Richtungen können durch die Nutzung mehrerer Mikrofone, wie in Abbildung 1 dargestellt, reduziert werden. So wie der Mensch in der Lage ist sich auf eine bestimmte Raumrichtung durch die Ausnutzung der Informationen beider Ohren zu konzentrieren und dadurch akustische Quellen aus anderen Richtungen zu unterdrücken, ist dies auch mit mehrkanaligen Signalverarbeitungskonzepten möglich [9][19][20].

Ein weiteres Konzept in der ASR ist es, viele Mikrofone im Raum zu verteilen und dann jenes auszuwählen, welches die beste Signalqualität für den Erkennungsprozess bietet [21]. Typischerweise ist dies das Mikrofon, welches sich am dichtesten zum Sprecher befindet, da es wahrscheinlich das beste Signal-zu-Rauschverhältnis bietet. Dennoch ist die Nutzung komplexerer Ansätze lohnenswert.

3.2 Konzepte in der automatischen Spracherkennung zur Verbesserung der Erkennerleistung für akustisch schwierige Bedingungen

Neben einer geeigneten Signalvorverarbeitung existieren weitere Konzepte, um die ASR Erkennerleistung für akustisch schwierige Bedingungen zu verbessern. Zwei wichtige Konzepte im Bezug auf die ASR-Algorithmen selbst werden im folgenden Abschnitt dargestellt.

Das erste Konzept ist es die Trainingsaufnahmen unter den gleichen Bedingungen aufzunehmen, wie in denen das ASR System später eingesetzt werden soll. Dies ermöglicht, dass die verwendeten akustischen Modelle, wie z. B. das Hidden Markov Modell (HMM), die Rauschcharakteristiken lernen kann, welche das Sprachsignal stören [8] [22]-[23][25]. Auch Faltungsrauschen wie z. B. die Charakteristik des Übertragungskanals und Raumhall können durch die akustischen Modelle bis zu einem gewissen Grad erlernt werden. Sobald sich allerdings die akustische Umgebung ändert (z. B. die Menge der Umgebungsgeräusche, die spektrale Rauschcharakteristik oder die Nachhallzeit) können die Vorteile eines an die erwarteten akustischen Bedingungen angepassten Trainings reduziert oder sogar ganz verloren gehen [8][23]. Daher ist diese Methode nur praktikabel, wenn die akustischen Bedingungen in denen das ASR System betrieben werden soll genau bekannt sind, um eine zuverlässige Verbesserung zu erhalten. Für den Fall, dass ein Spracherkenner in sich ändernden akustischen Umgebungen eingesetzt werden soll, welche aber vorweg bekannt sind, dann kann zwischen verschiedenen akustischen Modellen hin und her geschaltet werden, die mit den unterschiedlichen Störcharakteristiken trainiert wurden. Für diesen Fall ist jedoch eine zuverlässige Detektion der aktuellen akustischen Umgebungscharakteristik nötig, was ebenfalls ein schwieriges Problem darstellen kann [24].

Ein weiteres Konzept, um die Erkennungsraten unter akustisch schwierigen Bedingungen zu verbessern, ist eine geeignete Auswahl der akustischen Merkmale zu treffen, die verwendet werden, um die Eigenschaften von Sprache abzubilden. Diese müssen zwei wichtige Anforderungen erfüllen. Erstens müssen diejenigen Merkmale in einem Sprachsignal, die eine Unterscheidung zwischen den verschiedenen bedeutungsunterscheidenden Einheiten von Sprache erlauben, möglichst gut abgebildet werden. Und zweitens sollten Störungen wie Lärm und Nachhall diese Sprachmerkmale in der gewählten Darstellung möglichst wenig beeinflussen bzw. verdecken können.

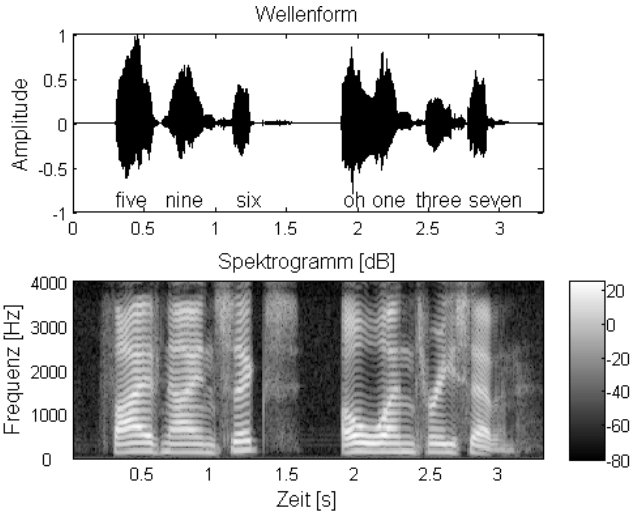

Abbildung 3: Zeitbereichsdarstellung einer Sprachsequenz (oberes Feld) und dessen spektro-temporale Darstelung bzw. Spektrogramm (unteres Feld).

Die meisten Merkmalsextraktionsverfahren in der ASR basieren auf einer spektralen Repräsentation der Wellenform des Sprachsignals. Um diese zu erhalten wird das akustische Signal zunächst in kurze überlappende Segmente von 20 bis 30 ms Länge aufgeteilt, die dann durch eine Fourier Transformation in den Frequenzbereich überführt werden. Dieses Vorgehen wird im Allgemeinen „short-time Fourier transformation“ (STFT) genannt. Das Ergebnis ist eine Zeit-Frequenz Darstellung, welche als Spektro-

ISBN 978-3-8007-3323-1

gramm bezeichnet wird (dargestellt im unteren Feld von Abbildung 3).
Typischerweise wird zusätzlich eine Filterbank, wie z. B. die Mel- oder Bark-Filterbank, welche durch das menschliche auditorische System motiviert ist, auf das Spektrogramm angewendet [8][22][23]. Diese Art Filterbank gruppiert bestimmte Frequenzabschnitte zu Bändern, wodurch die Frequenzauflösung des inneren menschlichen Gehörs approximiert werden soll. Für heutige Merkmalsextraktionsverfahren sind weitere Rechenschritte notwendig. Eine detaillierte Beschreibung der verschiedenen Verfahren wäre an dieser Stelle jedoch nicht im Sinne dieses Beitrags. Der interessierte Leser wird daher auf [8][22][23] verwiesen.

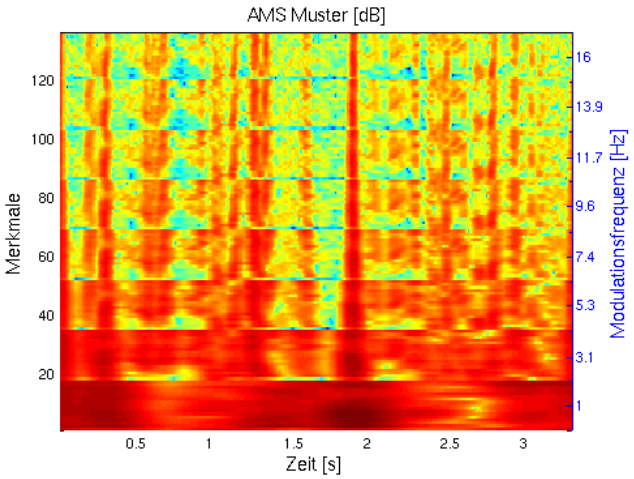

Abbildung 4: AMS Muster des in Abbildung 3 dargestellen Signals. Um alle 3 Dimensionen darstellen zu können, sind die Zeit versus Frequenz Darstellungen für jede einzelne Modulationsfrequenz übereinander gestapelt abgebildet.

Stattdessen wird im Folgenden ein Merkmalsextraktionsverfahren exemplarisch vorgestellt, um zu zeigen, welche Art von Ansatz in der aktuellen Forschung zur Verbesserung der Robustheit eines ASR System verfolgt wird. Aktuelle Forschungsansätze im Bereich der Merkmalsextraktion tendieren dazu längere Zeit-Trajektorien der spektralen Einhüllenden, die aus dem Spektrogramm gewonnen werden, zu analysieren. Als ein Beispiel wird hier das Amplitudenmodulationsspektrogramm (AMS) vorgestellt [26]. Dieser Merkmalstyp analysiert 300 bis 350 ms lange Zeit-Trajektorien aus dem Spektrogramm, welches vorher halbwellengleichgerichtet, quadriert und in Bark Bändern zerlegt wurde [27]. Dadurch werden die Amplitudenmodulationen des akustischen Signals analysiert. Anschließend werden die Modulationsfrequenzen mit einem Bandpassfilter eingegrenzt, so dass Modulationsfrequenzen kleiner als 1 Hz und größer als 16 Hz ignoriert werden. Damit werden nur die für Sprache relevanten Modulationsfrequenzen durchgelassen, wodurch andere Einflüsse außer Sprache bereits unterdrückt werden können [28]-[30].
Abbildung 4 zeigt ein Beispiel eines AMS Musters für das Signal aus Abbildung 3. Weil das AMS eine dreidimensionale Darstellung eines Signals ist, nämlich Zeit versus akustischer Frequenz und Modulationsfrequenz, ist es nicht leicht eine einfach interpretierbare graphische Darstellung zu erzeugen. Aus diesem Grund werden die Modulationsfrequenzen und die akustischen Frequenzen in Abbildung 4 gestapelt dargestellt. Dies ist zu erkennen durch die Achsenbeschriftung auf der rechten Seite, welche andeutet, dass die Zeit-Frequenz Muster für jede Modulationsfrequenz übereinander gestapelt sind. Für das Beispiel in Abbildung 4 resultiert dies in einem 136 dimensionalen Merkmalsvektor (17 Bark Bänder mal 8 Modulationsfrequenzen) pro Zeitabschnitt.
Es sollte erwähnt werden, dass dimensionsreduzierende Methoden, sowie bspw. eine Hauptkomponentenanalyse (principal component analysis, PCA), typischerweise verwendet werden, bevor diese Darstellung als akustisches Merkmal zur Klassifikation verwendet wird [8][25], da die Klassifikationsalgorithmen, wie z. B. HMMs, effizienter arbeiten, wenn die Dimension der Merkmale gering ist und die einzelnen Merkmalskomponenten dekorreliert sind [8][25].

4 Ein automatisches Spracherkennungsexperiment in einem AAL Szenario

In diesem Kapitel wird ein ASR Experiment vorgestellt, um den Einfluss von Raumhall und des Übertragungskanals auf die Leistung eines Spracherkenners zu demonstrieren. Dazu wurden Aufnahmen von acht männlichen Sprechern in einem AAL Living Lab gesammelt. Dieser Raum ist wie ein Wohnzimmer eingerichtet und mit fast unsichtbaren Deckenmikrofonen ausgerüstet. Die Deckenmikrofone repräsentieren eine Freisprecheinrichtung, die für das ASR Experiment verwendet wurde. Zusätzlich wurde ein Mikrofon direkt am Mund des Sprechers platziert. Für die Aufnahmen saßen die Sprecher auf einem Sessel in der Mitte des Raumes mit dem Kopf Richtung Fernseher gerichtet (vgl. Abbildung 5).
Die synchrone Aufzeichnung der gesprochenen Kommandos mit dem “close-talk” Mikrofon und mit einem der Deckenmikrofone, welches sich in der Mitte des Raumes befand (vgl. Abbildung 5), wurden dann genutzt um einen Spracherkenner zu trainieren. 25 verschiedene Schlüsselwörter wurden hierfür aufgenommen, um einen Erinnerungsassistenten (Kalendersystem) zu steuern. Jedes dieser Wörter wurde insgesamt 10-mal von jedem Sprecher in einer ruhigen Umgebung aufgezeichnet. Die Tests wurden mit dem sogenannten Kreuzvalidierungsverfahren durchgeführt [25]. D.h., dass die Aufnahmen von 7 der 8 Sprecher für das Training verwendet wurden und die Aufnahmen des übrigen Sprechers zum Testen. Alle 8 Kombinationsmöglichkeiten wurden auf diese Weise getestet und eine durchschnittliche Wortfehlerrate ermittelt.
Die in der ASR am häufigsten verwendeten akustischen Merkmale, nämlich die Mel-Frequency Cepstral Coefficients (MFCCs - hier inklusive des 0ten Cepstralen Koef-

ISBN 978-3-8007-3323-1

fizienten plus den dynamischen Merkmalen der Delta und Doppel-Delta Koeffizienten) [8][22][23], wurden auch für dieses Experiment ausgewählt. Der verwendete Klassifizierer beruht auf der statistischen Beschreibung durch lineare ganzwortbasierte HMMs mit jeweils 14 Zuständen (inklusive der zwei nicht emittierenden Zustände).

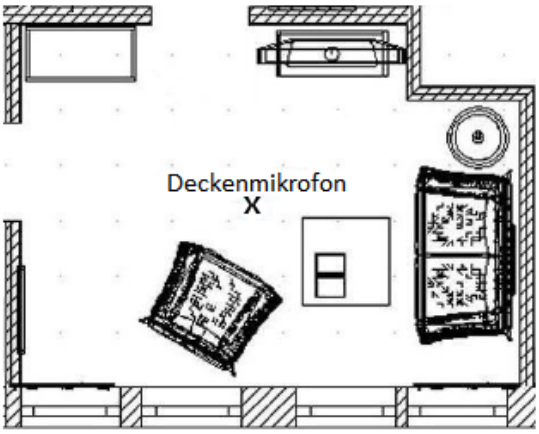

Abbildung 5: Grundriss der Umgebung im ALL Living Lab. Die Position des Deckenmikrofons, welches für das Experiment verwendet wurde, ist mit einem „X“ markiert. Der Sessel in der unteren Bildmitte zeigt die Position der Sprecher an.

Drei Testszenarien wurden definiert:

a) Die Aufnahmen des “close-talk” Mikrofons werden für das Training verwendet und die des Deckenmikrofons zum Testen.
b) Das Deckenmikrofon wird sowohl für das Training als auch zum Testen verwendet.
c) Das “close-talk” Mikrofon wird sowohl für das Training als auch zum Testen verwendet.

Table 1: ASR Ergebnisse für die verschiedenen Testszenarios. SD: Standardabweichung („Standard Deviation“).

Testszenario	**(a)**	**(b)**	**(c)**
WER (SD) in %	56,9 (26,5)	2,2 (2,3)	1,1 (2,1)

Tabelle 1 zeigt die Ergebnisse in Form von Wortfehlerraten (word error rates, WERs). Die WER wird berechnet aus dem Verhältnis von der Summe aller Fehler durch die Anzahl der gesprochnen Wörter. Fehler die hierbei auftreten können sind das Löschen (*DEL*), das Ersetzen (*SUB*) und das Einfügen eines Wortes.

$$WER = \frac{\#DEL + \#SUB + \#INS}{N} \qquad (1)$$

N entspricht hierbei der Anzahl der gesprochenen Wörter.
Die Ergebnisse zeigen, dass die Leistung des Spracherkenners sehr stark von den verwendeten Trainingsdaten abhängt. In dem Fall, dass die Aufnahmen für das Training und zum Testen unter den gleichen Umständen erzeugt wurden (siehe Ergebnisse der Testszenarios (b) and (c)) können sehr geringe WERs erreicht werden. Sobald die Aufnahmen des „close-talk“ Mikrofons zum Training verwendet werden, welche als frei von Raumhall betrachtet werden können, und hallige Sprache die durch die Deckenmikrofone aufgenommen wurde zum Testen verwendet wird, steigen die beobachtet WERs stark an (siehe Ergebnisse von (a)). Neben Raumhall spielen hier jedoch auch Verzerrungen des Kanals und internes Rauschen durch die Verwendung verschiedener Mikrofone für die Decke und für „close-talk“ eine wesentliche Rolle und führen zusätzlich zu den beobachteten schlechten WERs des Testszenarios (a).
Die Ergebnisse zeigen, dass es notwendig ist die spezifischen Umgebungsbedingungen und bekannte Störungen bereits bei der Vorbereitung der Trainingsdaten für ein ASR System zu berücksichtigen. Dies ist ein besonders effektives Konzept, wenn Freisprecheinrichtungen genutzt werden die fest in einem Raum verbaut sind, da Änderungen der räumlichen Umgebung hier nicht zu erwarten sind. Allerdings sollte an dieser Stelle erwähnt werden, dass zusätzlich auch andere Schwierigkeiten, die nicht durch dieses Experiment berücksichtigt wurden, verstärkt auftreten, wenn Freisprecheinrichtungen für die ASR genutzt werden. Diese Probleme ergeben sich aus der Tatsache, dass Freisprecheinrichtungen nicht nur die Sprache des gewünschten Sprechers erfassen, sondern auch andere Störquellen in der Umgebung, sowie z.B. andere sich unterhaltende Personen.

5 Zusammenfassung und Fazit

In diesem Beitrag wird ein Überblick über aktuelle Herausforderungen und Anforderungen für akustische Benutzerschnittstellen im AAL Kontext mit Fokus auf der automatischen Spracherkennung gegeben. Die diskutierten Ergebnisse werden zusätzlich durch ein ASR Experiment in einem AAL Living Lab unterstützt.
In Abschnitt 2 wurde eine Nutzerstudie präsentiert, in der die Akzeptanz einer sprachgesteuerten Kalenderanwendung evaluiert wurde. Außerdem wurde in der vorgestellten Studie die Toleranzgrenze von Probanden im Alter zwischen 63 und 75 Jahren für fehlerhafte Erkennungen untersucht. Für dieses Experiment wurde ein Mock-Up System verwendet, welches das ASR System simulierte. Das Ergebnis war, dass eine Sprachsteuerung von den älteren Nutzern akzeptiert wurde, auch wenn das System Erkennungsfehler produziert.
In Abschnitt 3 wurde ein Entwurf eines persönlichen Aktivitäts- und Haushaltsassistenten dargestellt. Basierend auf diesem beispielhaften Anwenderszenario wurden mögliche Herausforderungen an das ASR System diskutiert, welche typischerweise auftreten, wenn ASR zusammen mit einer Freisprecheinrichtung zur akustischen Interaktion mit assistiven Technologien genutzt wird. Zusätzlich wurden in diesem Zusammenhang verschiedene Konzepte zur Verbesserung der Robustheit eines ambieten sprachgesteuerten Gerätes vorgestellt. Außerdem wurde ein kleiner Einblick in die akustische Merkmalsextraktion für die ASR

ISBN 978-3-8007-3323-1

gegeben, um ein Beispiel für ein aktuelles Forschungsfeld in der ASR aufzuzeigen.
In Abschnitt 4 wurden schließlich Ergebnisse eines ASR Experiments mit einer Freisprecheinrichtung präsentiert, um die Bedeutung einer Berücksichtigung der spezifischen Bedingungen, in denen ein ASR System eingesetzt werden soll, zu verdeutlichen. Es konnte gezeigt werden, dass sich die Erkennerleistung eines Spracherkenners entscheidend verschlechtert, falls bekannte Störungen nicht im Training berücksichtigt werden. Die betrachteten Störungen in diesem Experiment sind Raumhall und die Mikrofoncharakteristik, die bei der Verwendung verschiedener Mikrofone zum Tragen kommt.
Als eine allgemeine Schlussfolgerung kann zusammengefasst werden, dass mit der ASR als eine Technologie zur Nutzerinteraktion mit assistiven Technologien bisher Teilziele erreicht wurden. Die Leistung eines modernen ASR Systems ist allerdings noch weit von der Leistungsfähigkeit eines Menschen entfernt. Dies gilt insbesondere dann, wenn die Spracherkennung zusammen mit entfernten Raummikrofonen (d. h. ambiente Mikrofonie) in beliebigen akustischen Umgebungen angewendet werden soll.
Da die Verwendung von ASR in Kombination mit Freisprecheinrichtungen eine sehr natürliche Art der Interaktion mit assistiven Technologien darstellt und weil die ambiente ASR großes Potential birgt, um assistive Technologien unaufdringlich zu machen und gleichzeitig einfach zu bedienen, sind Konzepte zur Verbesserung der Robustheit (wie in diesem Beitrag vorgestellt) aus technologischer Sicht, sowie aus der AAL anwendungsorientierten Sicht, sehr vielversprechend.

6 Danksagung

Diese Arbeit wurde in Teilen durch das niedersächsische Kultur- und Wissenschaftsministerium im Rahmen des Förderprogramms „Niedersächsisches Vorab“ innerhalb des niedersächsischen Forscherverbundes „Gestaltung altersgerechter Lebenswelten (GAL)“ gefördert.

7 Literatur

[1] European Commission Staff: Working Document. Europes Demografic Future: Facts and Figures. Commission of the European Communities. Mai, 2007.

[2] Statistisches Bundesamt: Demografischer Wandel in Deutschland – Heft 1 – Bevölkerungs- und Haushaltsentwicklung im Bund und in den Ländern. Dez, 2007.

[3] Statistisches Bundesamt: Demografischer Wandel in Deutschland – Heft 1 – Auswirkungen auf Krankenhausbehandlungen und Pflegebedürftige im Bund und in den Ländern. 2008.

[4] European Ambient Assisted Living Innovation Alliance: Ambient Assisted Living Roadmap. VDI/ VDE-IT AALIANCE Office, 2009.

[5] Alexandersson, J.; Zimmermann, G.; Bund J.: User interfaces for AAL: How can I satisfy all users? 2. Deutscher AAL Kongress, Berlin: Deutschland, 2009.

[6] Rennies, J.; Goetze, S.; Appell J.-E.: Personalized acoustic interfaces for human-computer interaction. In Human-Centered Design of E-Health Technologies: Concepts, Methods and Applications, Ziefle M. and Röcker C., Eds. IGI Global, 2010.

[7] Meyer, E. M.; Heuten, W.; Meis, M.; Boll, S.: Multimodal Presentation of Ambient Reminders for Older Adults, 3. Deutscher AAL Kongress. Berlin: Deutschland, 2010.

[8] Wölfel, M.; McDonough, J.: Distant Speech Recognition. Wiley, 2009. ISBN 978-0-470-51704-8.

[9] Mildner, V.; Goetze, S.; Kammeyer, K.-D.: Multi-Channel Noise-Reduction-Systems for Speaker Identification in an Automotive Acoustic Environment. In Proc. Audio Engineering Society (AES), 120th Convention, Paris: Frankreich, 20. - 23. Mai, 2006.

[10] Lippmann, R.: Speech recognition by machines and humans. J. Speech Communication, 22:1-15, 1997.

[11] Goetze, S.; Moritz, N.; Appell, J.-E.; Meis, M.; Bartsch, C.; Bitzer, J.: Acoustic User Interfaces for Ambient Assisted Living Technologies. Informatics for Health and Social Care 35(4), pp161-179, Dez., 2010.

[12] Meis, M.; Fleuren, T.; Meyer, E. M.; Heuten, W.: User centred design process of the personal activity and household assistant: Methodology and first results. 3. Deutscher AAL Kongress. Berlin: Deutschland, Jan. 2009.

[13] Schröder, J.; Wabnik, S.; van Hengel, P.W.J.; Goetze, S.: Detection and Classification od Acoustic Events for In-Home Care. In Proc. 4. Deutscher AAL Kongress. Berlin: Deutschland, 2011.

[14] Breining, C.; Dreiseitel, P.; Hänsler, E.; Mader, A.; Nitsch, B.; Puder, H.; Schertler, T.; Schmidt, G.;Tilp, J.: Acoustic Echo Control – An Application of Very-High-Order Adaptive Filters. IEEE Signal Processing Magazine, pp. 42–69, July 1999.

[15] Hänsler, E; Schmidt, G.: Speech and Audio Processing in Adverse Environments. Springer, 2008.

[16] Goetze, S.; Xiong, F.; Rennies, J.; Rohdenburg, T.; Appell, J.-E.: Hands-Free Telecommunication for Elderly Persons Suffering from Hearing Deficiencies. Proc. 12th IEEE International Conference on E-Health Networking, Application and Services (Healthcom’10); 2010 Jul 1-3; Lyon: Frankreich, 2010.

[17] Benesty, J; Morgan, D. R.; Sondhi, M. M.: A Better Understanding and an Improved Solution to the Specific Problems of Stereophonic Acoustic Echo Cancellation. IEEE Trans. on Speech and Audio Processing, vol. 6, no. 2, pp. 156–165, Mar 1998.

[18] Goetze, S.; Kallinger, M.; Kammeyer, K.-D.; Mertins, A.: Enhanced Partitioned Residual Echo Estimation, Proc. Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2006.

[19] Bitzer, J.; Simmer, K.U.: Superdirective microphone arrays. In M. S. Brandstein & D. Ward (Eds.), Microphone Arrays: Signal Processing Techniques and Applications. Berlin: Springer; pp. 19-38, 2001.

[20] Doblinger, G.: Localization and Tracking of Acoustical Sources. In Topics in Acoustic Echo and Noise Control. Berlin: Springer, pp. 91-122, 2006.

[21] Wolf, M.; Nadeu, C.: On the potential of channel selection for recognition of reverberated speech with multiple microphones. Interspeech, Japan, 2010.

[22] Benesty, J.; Sondh, M. M.; Huang, Y.: Springer Handbook of Speech Recognition. Springer: New York, 2008.

[23] Rabiner, L.; Juang, B-H.: Fundamentals of speech recognition. Prentice-Hall, 1993.

[24] Tchorz J, Kollmeier B: Automatic classification of acoustical situation using amplitude modulation spectrograms. J. Acoust. Soc. Am. 105 (2), 1157. 1999.

[25] Bishop C. M.: Pattern recognition and machine learning. Springer, 2006.

[26] Kollmeier, B.; Koch, R.: Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. J. Acoust Soc Am 95(3), pp. 1593-1602, 1994.

[27] Moritz, N.; Meyer, B. T.; Anemüller, J., Kollmeier, B.: Robustheit automatischer Spracherkennung mit Amplitudenmodulationsspektrogrammen. 36. Jahrestagung für Akustik (DAGA), Berlin: Deutschland, 2010.

[28] Kanedera, N.; Arai, K.; Hermansky, H.; Pavel, M.: On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication 28, pp. 43-55, 1999.

[29] Kanedera, N.; Hermansky, H.; Arai, T.: On properties of modulation spectrum for robust automatic speech recognition. Proc. ICASSP 1998, pp. 613-616, 1998.

[30] Hermansky, H.: RASTA processing of speech. IEEE Trans. Speech and Audio Processing, 2(4), pp. 578-589, 1994.

ISBN 978-3-8007-3323-1