UNTERSUCHUNGEN ZUR GÜTE DER SIMULATION EINER SPRACHEINGABE IM FREISPRECHMODUS BEI DER EVALUIERUNG VON SPRACHERKENNUNGSSYTEMEN

Andreas Kitzig und Hans-Günter Hirsch

iPattern, Institut für Mustererkennung, Hochschule Niederrhein, 47805 Krefeld
andreas.kitzig@hs-niederrhein.de

Abstract: Bei der Entwicklung von robusten Spracherkennungssystemen ist es von großem Interesse, zur Evaluation der Leistungsfähigkeit eines Systems Sprachsignale zur Verfügung zu haben, die möglichst realistisch die akustischen Bedingungen praxisrelevanter Störszenarien beinhalten. Durch die Evaluation wird gewährleistet, dass das System im späteren Praxiseinsatz zuverlässig funktioniert und die bestmögliche Erkennungsrate liefert. Eine Möglichkeit zur Generierung solcher Sprachdaten besteht in der Simulation der akustischen Bedingungen, z.B. durch die additive Überlagerung von ungestörten Sprachsignalen und Störgeräuschen oder einer Faltung mit geeigneten Raumimpulsantworten zur Simulation einer Spracheingabe im Freisprechmodus. Bei einer Simulation der akustischen Bedingungen stellt sich jedoch die Frage, wie gut die Simulation die reale Aufnahme von Sprachsignalen in der jeweiligen akustischen Umgebung widerspiegelt. Dies wird im Rahmen der hier vorgestellten Arbeiten für eine Spracheingabe im Freisprechmodus in Räumen untersucht. Dazu wurden mittels eines eigenen Aufnahmeaufbaus Sprachdaten der TiDigits Sprachdatenbank [3] in insgesamt sechs verschiedenen Räumen wiedergeben und an 19 unterschiedlichen Aufnahmepositionen aufgezeichnet, um reale Sprachdaten im Freisprechmodus zu erzeugen. Zusätzlich wurde in jeder Aufnahmeposition die Raumimpulsantwort bestimmt und ein entsprechender Datensatz künstlich verhallter Daten generiert. Die Güte der realen und der simulierten Daten wurde abschließend anhand von verschiedenen Spracherkennungsexperimenten untersucht. Der vorliegende Text ist wie folgt aufgebaut: Nach einer Einleitung, in der die Simulation von geeigneten Testdaten für Spracherkennungssysteme theoretisch betrachtet wird, folgt ein Überblick über die praktische Umsetzung zur Erzeugung der realen und simulierten Testdaten. Anschließend werden die für die Bestimmung der Güte der generierten Testdaten verwendeten Spracherkennungssysteme dargestellt. Abschließend erfolgt die Darstellung und Diskussion der Ergebnisse.

1 Simulation von Testdaten zur Evaluation von Spracherkennungssystemen

Bei der Entwicklung von robusten Spracherkennungssystemen ist es von großer Bedeutung, dass neu entwickelte Systeme mit Daten getestet werden, die reale Störszenarien, welche bei dem Einsatz eines Spracherkennungssystems auftreten können, bestmöglich nachbilden. So wird garantiert, dass die entwickelten Systeme im späteren Einsatz, z.B. in einem KFZ, eine nahezu identische Leistungsfähigkeit im Vergleich zum Labor-Testbetrieb aufweisen. Zur Erzeugung von verhallten Testdaten, die möglichst realistisch sind, wird im Allgemeinen eine Faltung der ungestörten Sprachdaten mit einer zuvor bestimmten Raumimpulsantwort $h_{RIA}(t)$ vorgenommen.

$$y_{convolved}(t) = s(t) * h_{RIA}(t)$$

Durch diese Maßnahme kann ein nahezu realistisches Abbild einer räumlichen Umgebung erzeugt werden. Um präzisere Daten für die Evaluation von Erkennungssystemen nutzen zu können, empfiehlt es sich, Daten direkt in einer entsprechenden realen Umgebung aufzuzeichnen. Diese enthalten im Vergleich zu den simulierten Daten eine noch präzisere Beschreibung der Raumcharakteristik $h_{Raum}(t)$, Anteile von Störeinflüssen des verwendeten Aufnahme-Equipments $n_{Equ}(t)$ sowie Anteile der eventuell vorhandenen Störumgebung im Raum $n_{Umg}(t)$:

$$y_{recorded}(t) = \left(s(t) * h_{Raum}(t)\right) + n_{Equ}(t) + n_{Umg}(t)$$

Ein großer Nachteil der realen Daten ist jedoch die zeit- und personalaufwendige und somit kostenintensive Erzeugung. Daher stellt sich die Frage, ob die Ergebnisse von Erkennungsexperimenten zur Evaluation der Leistungsfähigkeit von Erkennungssystemen überhaupt eine höhere Aussagekraft besitzen, wenn reale Daten anstatt der simulierten Daten verwendet werden. Dieser Punkt wird in der vorliegenden Arbeit untersucht. Dazu werden reale und simulierte Testdaten im Rahmen von Erkennungsexperimenten miteinander verglichen.

Da die meisten Sprachdatenbanken entweder aus real aufgezeichneten Daten oder aus simulierten Daten bestehen, wurde zur Durchführung der Tests eine eigene Sprachdatensammlung erstellt, die sich sowohl aus simulierten als auch aus realen Testdaten zusammensetzt. Die Sprachdatensammlung beinhaltet in dem jeweiligen Testset 8700 Äußerungen (28583 Ziffern / ca. 4,5h Material) von 56 männlichen und 57 weiblichen Sprechern. Als Basisdaten wurde die englischsprachige TiDigits Datenbank verwendet.

Zur Erzeugung der realen Testdaten wurden zunächst mit einem eigenen Aufnahmesystem reale Aufnahmen von Sprachsignalen in unterschiedlichen räumlichen Umgebungen erstellt. Dazu wurden die Basisdaten in sechs verschiedenen Räumen wiedergegeben und an 19 Positionen aufgezeichnet. Zusätzlich wurde bei jeder Aufnahmeanordnung die Raumimpulsantwort (RIA) geschätzt, aus der anschließend die simulierten Daten generiert werden können.

2 Erstellung der Testdaten

Das zur Erstellung der Testdaten verwendete Aufnahmesystem besteht aus mobilen netzunabhängigen Vorverstärkern und Mikrofonen, einem Notebook und einem Aktivlautsprecher. Der prinzipielle Aufnahmeaufbau ist in Abbildung 1 dargestellt. Über den Aktivlautsprecher werden englische Ziffern und Ziffernketten der TiDigits Sprachdatensammlung in den verschiedenen Räumen wiedergegeben und unter Verwendung des Mikrofons M2 aufgezeichnet. Bei den Aufnahmen wurden alltägliche Freisprechszenarien in den Räumen der Hochschule Niederrhein nachgestellt, wobei pro Raum an zwei bis vier unterschiedlichen Positionen Daten aufgezeichnet wurden. Die Aufnahmen wurden in drei Besprechungsräumen, zwei Büros und einem Laborraum durchgeführt.

Zusätzlich wurden zu den realen Aufnahmen noch die raumakustischen Eigenschaften für die 19 Aufnahmepositionen bestimmt. Dazu wurde der Aufnahmeaufbau leicht variiert. Der Aufbau zur Bestimmung der raumakustischen Eigenschaften ist zusätzlich in Abbildung 1 (rot gestrichelt) dargestellt. Anstelle des Notebooks wird nun ein CD-Player verwendet, um über den Lautsprecher eine Maximum-Length-Sequenz (MLS) wiederzugeben. Der Lautsprecher und das Mikrofon M2 befinden sich weiterhin an den ursprünglichen Positionen zur Aufnahme und Wiedergabe der Testdaten. Mit Hilfe des zusätzlichen Mikrofons M1 wird das vom Lautsprecher abgestrahlte Signal direkt aufgezeichnet. Dies ist notwendig, um die Übertragungscharakteristik des Lautsprechers bei der Berechnung der RIA zu kompensieren. Die beiden Mikrofonsignale werden synchron auf dem Notebook gespeichert. Anschließend wird mittels einer Software die Raumimpulsantwort daraus geschätzt [7].

Bei der Schätzung der RIA wird eine wichtige Eigenschaft der MLS ausgenutzt. Die Autokorrelationsfunktion der MLS ist näherungsweise ein Dirac-Impuls, besonders für lange ML Sequenzen. Das durch Mikrofon M2 aufgezeichnete Signal kann als Ergebnis der Faltung des Signals von Mikrofon M1 und der Raumimpulsantwort betrachtet werden. Somit kann die Kreuzkorrelierte der Mikrofonsignale M1 und M2 als Faltung der Autokorrelationsfunktion der MLS mit der Impulsantwort des Raumes betrachtet werden. Bedingt durch die Tatsache, dass die Autokorrelationsfunktion der MLS einem Dirac-Impuls entspricht, führt die Berechnung der Kreuzkorrelierten der Mikrofonsignale M1 und M2 direkt zu der gesuchten Raumimpulsantwort.

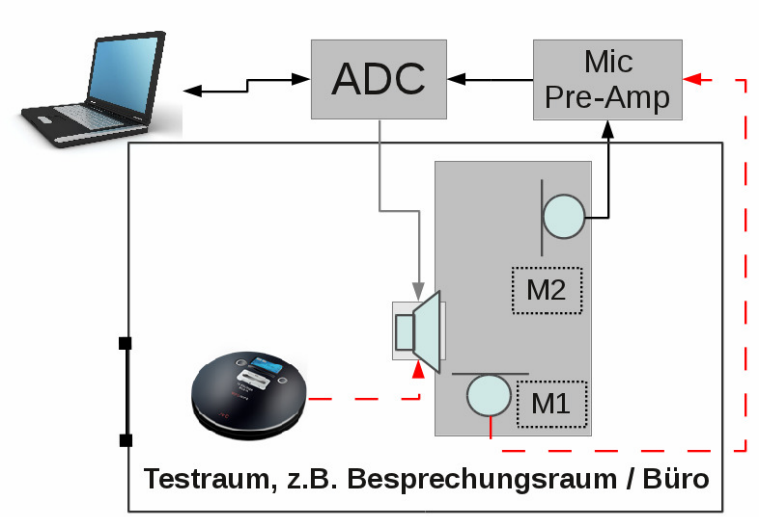

Abbildung 1 - Blockschaltbild des Aufnahmeaufbaus und des Aufbaus zu Bestimmung der Raumimpulsantwort

Anhand der geschätzten RIAs wird anschließend durch eine Faltung der ungestörten Sprachsignale mit der jeweiligen RIA eine simulierte Version $y_{convolved}(t)$ der aufgezeichneten Sprachsignale $y_{recorded}(t)$ erzeugt. Dies geschieht unter Verwendung eines eigenen Tools „Sireac“ [6].

Da die Daten $y_{recorded}(t)$ während des normalen Arbeits- und Vorlesungsbetriebs der Hochschule Niederrhein aufgezeichnet wurden, kann es vorkommen, dass die real aufgezeichneten Daten ein zusätzliches Störgeräusch $n_{Umg}(t)$ beinhalten, das sich in Form von leisen Stimmen und Trittgeräuschen auf den Fluren vor den Aufnahmeräumen darstellt. Prinzipiell wurde aber versucht, diese zusätzliche Störung weitestgehend zu vermeiden.

3 Erkennungssysteme

Um die Güte der simulierten Daten $y_{convolved}(t)$ mit den real aufgezeichneten Daten $y_{recorded}(t)$ vergleichen zu können, werden verschiedene Erkennungsexperimente durchgeführt. Anhand der Auswertung der resultierenden Erkennungsergebnisse wird die Güte der Simulation im Vergleich zu realen Daten beurteilt. Insgesamt werden drei robuste und ein nicht robustes Spracherkennungssystem zur Durchführung der Experimente eingesetzt. Zwei der drei Systeme zur robusten Spracherkennung wurden im Rahmen früherer Arbeiten entwickelt [1,2], bei dem dritten Verfahren handelt es sich um ein durch ETSI standardisiertes Verfahren [5].

Das durch ETSI standardisierte Verfahren wird mit dem Namen „ETSI 2“ referenziert. Bei diesem Verfahren werden in der Sprachanalyse robuste Merkmale extrahiert. Hierbei wird versucht, Merkmale zu gewinnen, die möglichst unabhängig von der akustischen Aufnahme-

situation sind. Dazu werden die Sprachdaten mittels eines zweistufigen Wiener-Filters im Zeitbereich gefiltert, um Hintergrundgeräusche zu minimieren. Die Filterkoeffizienten werden zuvor im Spektralbereich geschätzt und anschließend geglättet und in den Zeitbereich transformiert. Im nächsten Verarbeitungsschritt erfolgt ein SNR- abhängiges Waveform Processing des gefilterten Signals. Zur Berechnung von Cepstral-Koeffizienten wird das vorverarbeitete Signal in einem Verarbeitungsblock zur Merkmalsextraktion analysiert.

Zusätzlich ist ein Verarbeitungsblock zur „blinden" Kompensation unbekannter Frequenzgänge in die Verarbeitungskette integriert. Hierzu wird abschließend das durch die Cepstral-Koeffizienten beschriebene Spektrum mit einem „mittleren" Sprachspektrum bzw. den zugehörigen Cepstral-Koeffizienten verglichen.

Das zweite Verfahren „HGH robust" verwendet eine Sprachanalyse, bei der ebenfalls robuste akustische Merkmale gewonnen werden, die möglichst unabhängig von der akustischen Aufnahmesituation sind. Dazu wurde die Cepstralanalyse um eine adaptive Filterung im Spektralbereich ergänzt, mit der Hintergrundstörgeräusche reduziert werden sollen. Die adaptive Filterung beruht auf einer Schätzung des Spektrums der Hintergrundstörung und verwendet als speziellen Verarbeitungsschritt eine Glättung von Cepstralparametern in zeitlicher Richtung [4] zur Bestimmung der Filtercharakteristik.

Auch bei diesem Verfahren ist ein Verarbeitungsblock zur „blinden" Kompensation unbekannter Frequenzgänge integriert. Dieser orientiert sich bezüglich Funktionalität und Aufbau an der blinden Kompensation von „ETSI 2". Im Gegensatz zu dem Verfahren „ETSI 2" ist die Störgeräuschkompensation bei „HGH robust" in dem Verarbeitungsblock zur Analyse integriert und wird nicht separat vor der Analyse ausgeführt. Dadurch kann die für die Analyse benötigte Bearbeitungszeit verkürzt werden, indem z.B. unnötige Transformationen und Rücktransformation in den Frequenzbereich bzw. aus dem Frequenzbereich vermieden werden.

Das dritte Verfahren „HGH adapt" arbeitet mit einer Anpassung der akustischen Merkmale, die in den Referenzmustern enthalten sind, an die jeweilige Störsituation, an unbekannte Frequenzgänge und an den Hall des Raumes. Bei dem Adaptionsverfahren werden die Mittelwerte, die die Gauß-Verteilungen der akustischen Merkmale in den Zuständen eines statistisch basierten Hidden-Markov Modells festlegen, an die jeweiligen akustischen Bedingungen bei einer Spracheingabe angepasst. Dazu wird auch bei diesem Verfahren das Spektrum der Hintergrundstörung geschätzt. Zudem erfolgt eine Schätzung der Nachhallzeit, mit der grob der Einfluss eines Raumes im Fall einer Spracheingabe im Freisprechmodus beschrieben werden kann, sowie eine Schätzung von unbekannten Frequenzgängen.

Das nicht robuste System „HGH" verwendet eine einfache Cepstralanalyse zur Merkmalsextraktion. „HGH" wird als Referenzsystem verwendet, um eine minimal erreichbare Erkennungsrate (Baseline) zu ermitteln.

Bei allen vier Systemen werden als akustische Merkmale 12 Cepstralparameter und ein Energiekoeffizient sowie deren erste und zweite Ableitung, die so genannten Delta und Delta-Delta Koeffizienten, verwendet. Insgesamt gehen 39 Merkmalswerte aus der Analyse kurzer Sprachsegmente mit einer Dauer von ca. 25ms und in einem zeitlichen Abstand von 10ms hervor.

4 Erkennungsexperimente

In Abbildung 2 sind die bei den Erkennungsexperimenten zur Bestimmung der Güte erzielten Ergebnisse in Form von Wort-Erkennungsraten dargestellt. Bei den Experimenten wurden alle 8700 Testäußerungen für die simulierten Daten „convolved" und für die real aufgezeichneten Daten „recorded" verwendet. Zusätzlich sind die Ergebnisse eines Experiments mit den

ungestörten Basis-Daten „clean“, die zur Erzeugung der gestörten Daten verwendet wurden, angegeben. Diese stellen die maximal erreichbare Erkennungsrate der Systeme dar.

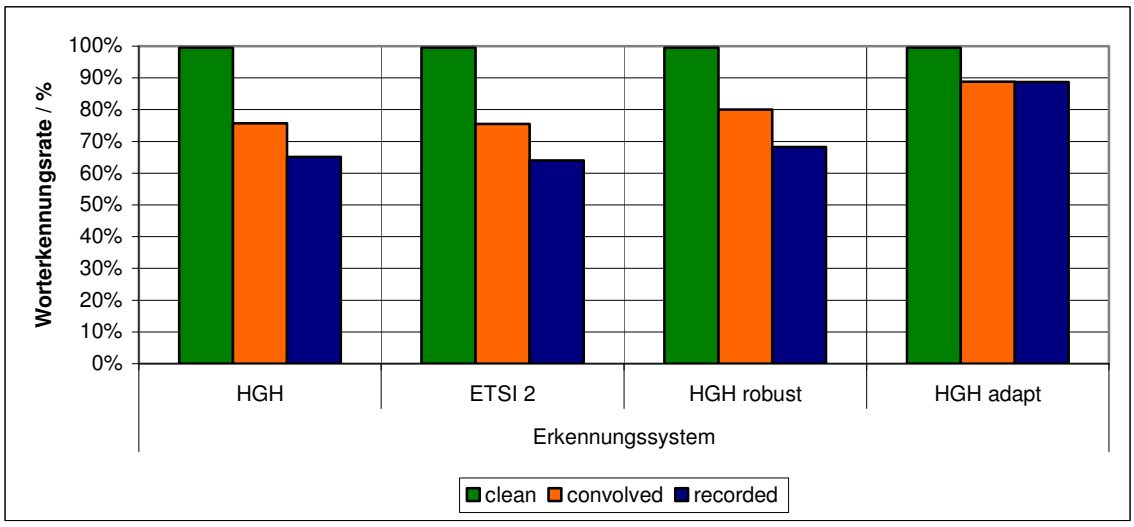

Abbildung 2 – Erkennungsergebnisse unter Verwendung aller 8700 Testäußerungen

Werden die Ergebnisse der gestörten Testumgebungen für „HGH“ verglichen, wird deutlich, dass sich für die realen Daten „recorded“ eine um etwa 11% schlechtere Erkennungsrate im Vergleich zu den simulierten Daten „convolved“ ergibt. Diese ist wahrscheinlich auf zusätzliche additive Störanteile in den realen Daten, bedingt durch das Aufnahme-Equipment und durch Störungen im Raum, zurückzuführen.

Das robuste System „ETSI 2“, das ein Verfahren zur Reduktion additiver Störanteile beinhaltet, liefert im Vergleich zu den Baseline-Ergebnissen von „HGH“ keine Verbesserung der Erkennungsrate. Speziell im Fall der realen Daten „recorded“, die neben dem Nachhall auch additive Störanteile beinhalten, sind die Erkennungsraten sogar geringfügig schlechter. Diese Verschlechterung der Erkennungsrate ist vermutlich auf die Schätzung der Filterkoeffizienten im Frequenzbereich für das Wiener-Filter zurückzuführen. Die Funktion zur Schätzung der Filterkoeffizienten ist auf die Verarbeitung von Daten, die mit additiven Störungen überlagert sind, ausgelegt. Werden hallige Daten verarbeitet, erfolgt möglicherweise auf Grund der „Verschmierung“ des Spektrums durch den Nachhall keine korrekte Schätzung der Filterkoeffizienten. Dies wirkt sich direkt auf die Störreduktion aus und führt zu einem Einbruch der Erkennungsrate.

Das robuste System „HGH robust“ liefert im Vergleich zu „ETSI 2“ eine Verbesserung der Erkennungsrate, jedoch liegt diese nur geringfügig über den Baseline Ergebnisse von „HGH“. Wie auch schon bei „ETSI 2“ vermutet wurde, liegt die geringe Verbesserung der Erkennungsrate bei „HGH robust“ vermutlich darin begründet, dass „HGH robust“ auf die Reduktion von additiven Störungen ausgelegt ist. Bei der Verarbeitung von halligen Daten kommt es bei „HGH robust“ zu einer Beeinträchtigung der Störschätzung in Folge des Nachhalls und zu einer fehlerbehafteten Schätzung der Filterkoeffizienten. Die nicht optimal geschätzten Filterkoeffizienten führen dazu, dass die anschließende Störreduktion nur eingeschränkt funktioniert. Somit resultiert lediglich eine relativ geringe Verbesserung der Erkennungsrate.

Das System „HGH adapt“ liefert die besten Ergebnisse, die sich zusätzlich für beide Testdatensätze „convolved“ und „recorded“ beinahe gleichen. Dies ist darauf zurückzuführen, dass die Referenzmodelle sowohl an den Nachhall als auch an die Störgeräusche in den Testdaten adaptiert werden können. Im Vergleich zu den Ergebnissen der „clean“ Daten wird jedoch deutlich, dass auch durch „HGH adapt“ keine vollständige Anpassung der Referenzmodelle an die Störsituation erreicht werden kann. Die Erkennungsrate wird im

Vergleich zu den anderen Systemen verbessert, jedoch wird die Erkennungsrate von über 99% im „clean“ - Fall nicht erreicht.

Die in Abbildung 2 vorgestellten Ergebnisse für die 8700 Testäußerungen sind über alle Räume und Aufnahmepositionen gemittelt. Um eine detailliertere Übersicht über die Erkennungsraten an den einzelnen Aufnahmeposition zu erhalten, sind nachfolgend die Ergebnisse für einen Besprechungsraum aus dem Aufnahme-Set nach Aufnahmepositionen getrennt dargestellt. Um einen besseren Überblick zu ermöglichen, ist in Abbildung 3 zunächst eine Skizze des Raumes sowie der Aufnahmepositionen mit den jeweiligen Abständen zwischen Lautsprecher und Mikrofon in dem Besprechungsraum (links) sowie ein Foto des Besprechungsraums (rechts) dargestellt. Die geschätzte Nachhallzeit T_{60} für den Besprechungsraum beträgt ca. 0,7s, der Raum hat Seitenmaße von ca. 13m x 6m.

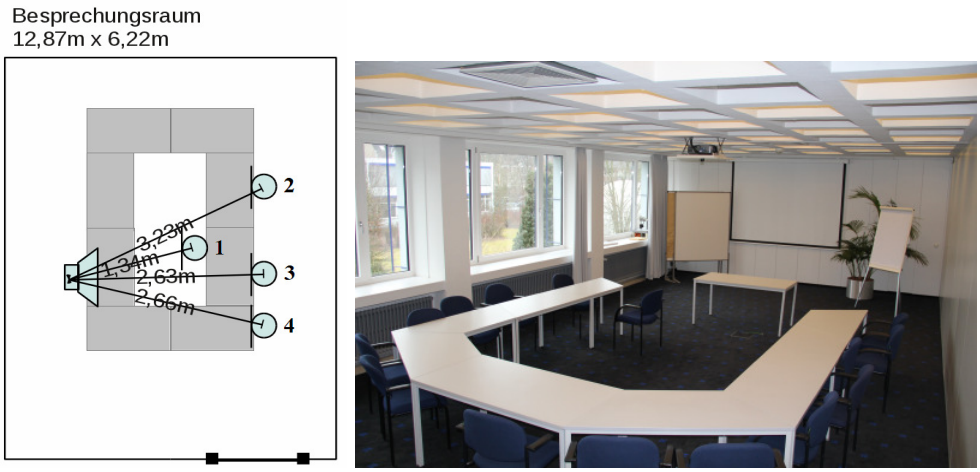

Abbildung 3 – Skizze der Aufnahmepositionen (links) und Foto des Besprechungsraums 8rechts)

Da die Ergebnisse des Basisexperiments unter Verwendung des nicht robusten Systems „HGH“ bei der Beurteilung der Güte der realen und simulierten Daten eine relativ objektive Beurteilung ermöglichen, da keinerlei Störunterdrückung etc. stattfindet, sind in Abbildung 4 nur die Ergebnisse des Systems „HGH“ dargestellt. Bei dem Vergleich der Ergebnisse der realen Testdaten mit den Ergebnissen der simulierten Testdaten in Abbildung 2 kann festgestellt werden, dass die „convolved“ Daten nicht vollständig die Störsituation in den realen Daten simulieren. Daher wurden zusätzliche simulierte Daten für die Experimente erzeugt, die neben den geschätzten Raumimpulsantworten auch einen additiven Störanteil $n_{Equ}(t)$ enthalten, der die Störeinflüsse des Aufnahmesystems beschreibt.

$$y_{\text{sim\_recorded}}(t) = \left(s(t) * h_{RIA}(t)\right) + n_{Equ}(t)$$

Dadurch können die simulierten Daten den realen Daten noch weiter angenähert werden. Das verwendete Störgeräusch wurde aus Aufnahmen in einem reflexionsarmen Raum unter Verwendung des vorgestellten Aufnahmeaufbaus extrahiert und beschreibt nur die Störeinflüsse des Equipments ohne akustische Merkmale des Testraumes. Die neuen Testdaten wurden mit einem Signal zu Rauschleistungsverhältnis (SNR) von ca. 20dB mit dem bereits vorgestellten Tool „Sireac“ erzeugt. Das eingestellte SNR entspricht in etwa dem mittleren SNR in den realen Daten. Die neu erzeugten Testdaten werden als „sim_recorded“ referenziert.

Die Ergebnisse für die Daten „recorded“ und „convolved“ zeigen erwartungsgemäß die gleichen Tendenzen im Vergleich zu den Ergebnissen in Abbildung 2. Bei Verwendung der realen Daten „recorded“ fallen die Erkennungsraten im Vergleich zu den simulierten Daten „convolved“ erheblich schlechter aus. Dies ist, wie bereits weiter oben erwähnt wurde,

teilweise auf zusätzliche Störanteile in den realen Daten zurückzuführen. Werden die Erkennungsraten für die einzelnen Positionen einer Testumgebung miteinander verglichen, zeigen die simulierten Daten „convolved“ erwartungsgemäß eine Abhängigkeit zwischen der Erkennungsrate und der Entfernung zwischen Quelle und Senke. Mit zunehmendem Abstand zwischen Lautsprecher- und Mikrofonposition kommt es zu einer Abnahme der Erkennungsrate auf Grund des zunehmenden Hallanteils. Dieser resultiert aus der Abnahme des Direktschallanteils in dem Verhältnis von Direktschall zu Raumschall bei zunehmendem Abstand.

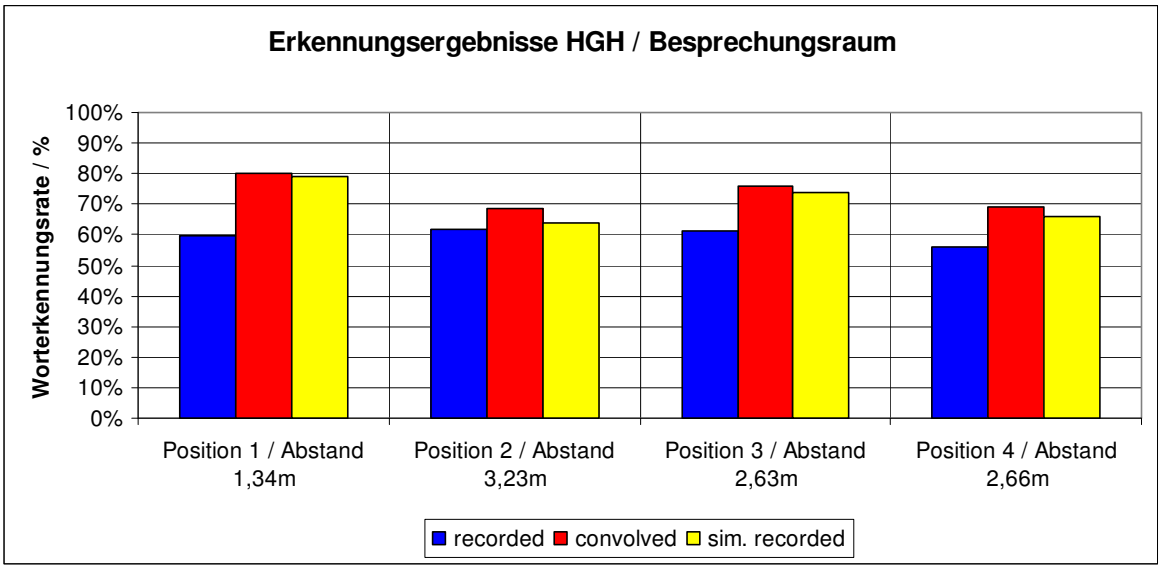

Abbildung 4 – Erkennungsergebnisse für einen einzelnen Testraum nach Positionen getrennt - Erkennungssystem HGH

Die Ergebnisse für die realen Daten zeigen diese Tendenz nicht und unterscheiden sich nur geringfügig für die einzelnen Positionen im Raum voneinander. Dies lässt darauf schließen, dass die additiven Störanteile in den realen Daten offenbar einen stärkeren Einfluss auf die Erkennungsleitung als die raumakustischen Anteile in den Daten besitzen. Aus diesem Grund zeigt sich bei einem Vergleich der Ergebnisse „recorded“ und „convolved“ erneut, dass durch die Simulation nicht die realen Daten nachgebildet werden können. Bei den Ergebnissen der simulierten Testdaten „sim_recorded“ zeigt sich erwartungsgemäß eine schlechtere Erkennungsrate im Vergleich zu „convolved“, bedingt durch die zusätzlichen Störanteile in den Testdaten. Jedoch können auch mit den Testdaten „sim_recorded“ nicht die Erkennungsraten bzw. die Tendenzen der Erkennungsraten der realen Daten „recorded“ erreicht werden.

Um die Ursache hierfür zu finden, werden in Tabelle 1 die Fehlerraten in der Position 4 in Einfügungen, Verwechslungen und Auslöschungen aufgeschlüsselt. Aus den Fehlerraten wird deutlich, dass die Anzahl der Auslöschungen in beiden Fällen ähnlich ist und auch die Anzahl der Ersetzungen bei den „recorded“ Daten nimmt nur geringfügig zu. Somit tragen diese Punkte kaum zu einer Reduktion der Erkennungsrate bei. Auffällig ist der um ca. 5,6 % gestiegene Anteil der Einfügungen bei den realen Daten. Dieser Anstieg liegt darin begründet, dass die realen Daten mit einer zusätzlichen Vor- und Nachlaufzeit aufgezeichnet wurden, um die Testäußerungen vollständig zu erfassen. Sie sind somit länger als die „convolved“ Daten. Die zusätzlichen Anteile vor und hinter den eigentlichen Sprachäußerungen enthalten jedoch meistens keine Sprache sondern nur Störanteile. Bei Erkennungsexperimenten werden mit hoher Wahrscheinlichkeit auf diese Bereiche fälschlicherweise Wortmodelle und kein Pausenmodell abgebildet, die bei der späteren Evaluation der Ergebnisse als Einfügung auftauchen. Dieser Effekt wird bei den simulierten Daten nicht modelliert, was zu der Abweichung der Ergebnisse von „recorded“ und „sim_recorded“ führt.

Testdaten	Auslöschungen	Ersetzungen	Einfügungen
sim_recorded	11,46 %	21,14 %	1,67 %
recorded	11,79 %	25,16 %	7,24 %

Tabelle 1 – Fehlerraten für die Testdaten an Position 4 / Abstand 2,66m

5 Fazit

Aus dem Vergleich der Ergebnisse der Experimente zur Bestimmung der Güte der Simulation geht hervor, dass Daten, die eine Spracheingabe im Freisprechmodus simulieren, eine geringere Güte im Vergleich zu realen Daten aufweisen. Selbst wenn zusätzliche additive Störanteile in den simulierten Daten enthalten sind, kann die reale Störsituation nur näherungsweise simuliert werden. Dies ist auf ein hohes Maß an Variabilität der additiven Stör-Komponente in den realen Daten zurückzuführen, die nur ungenau durch Störgeräuschproben modelliert werden kann. Bei den realen Daten kann es zusätzlich vorkommen, dass geringe Anteile von Umgebungsstörgeräuschen enthalten sind. Diese Punkte führen zu einer Beeinträchtigung der Erkennungsleistung bei den realen Daten und können bei den simulierten Daten nur bedingt berücksichtigt werden. Des Weiteren weisen die realen Daten eine größere zeitliche Länge als die simulierten Daten auf. Dies führt dazu, dass gerade in den langen Abschnitten vor und hinter der eigentlichen Sprachäußerung, die nur Störanteile aufweisen, Wortmodelle anstelle des Pausenmodells fälschlich erkannt werden. Hierdurch kommt es zu zusätzlichen Fehlern durch Einfügungen bei der Erkennung, die bei den simulierten Daten nicht auftreten. Dies erschwert den direkten Vergleich zwischen den Erkennungsraten der simulierten und der realen Daten und erklärt die relativ große Abweichung der Ergebnisse. Um die Erkennungsexperimente vergleichbarer zu gestalten, bietet es sich daher in weiteren Experimenten an, die „überschüssigen“ Anteile in den realen Daten, die nur Störungen enthalten, bei der Erkennung nicht zu berücksichtigen. Dazu könnte beispielsweise mittels einer „Voice Activity Detection“ (VAD) das Testsignal in Bereiche, die Sprache enthalten und Bereiche, die nur Hintergrundstörungen enthalten, aufgeteilt werden. Bei der anschließenden Erkennung werden nur die Anteile des Signals, welche Sprache enthalten, verarbeitet. Durch diese Maßnahme sollte es möglich sein, die Differenzen der Fehlerraten aus Experimenten mit realen und simulierten Daten in Tabelle 1 um etwa die Hälfte zu reduzieren. Insbesondere die Anzahl der Einfügungen sollte sich so minimieren lassen.

Literatur

[1] H.G. Hirsch. Automatic speech recognition in adverse acoustic conditions, in Advances in Digital Speech Transmission, John Wiley and sons, 2008

[2] H.G. Hirsch, A. Kitzig: Robust Speech Recognition by Combining a Robust Feature Extraction with an Adaptation of HMMs, 9. ITG Fachtagung Sprachkom., Okt. 2010

[3] R.G. Leonard, “A database for speaker independent digit recognition, ICASSP84, Vol.3, p.42.11, 1984

[4] C. Breithaupt, R. Martin, DFT based speech enhancement for robust automatic speech recognition, ITG Fachtagung Sprachkommunikation, Aachen, 2008

[5] ETSI ES 202050, “STQ; Distributed Speech Recognition, Advanced Front-End Feature Extraction Algorithm, Compression Algorithm”, ETSI ES 202 050 v1.1.3, Nov. 2003.

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

[7] H.G. Hirsch, A. Kitzig, K. Linhard: Simulation of the Hands-free Speech Input to Speech Recognition Systems by Measuring the Room Impulse Responses. 9. ITG Fachtagung Sprachkommunikation, Okt. 2010