Arbeitsbericht zum 4. Projektabschnitt

Andreas Kitzig

Juni 2010

Projektbericht zum Vorhaben „Robuste Spracherkennung in gestörter Umgebung durch die Kombination einer robusten Merkmalsextraktion und einer Adaption der Referenzmuster“, gefördert durch das BMBF im Rahmen des Förderprogramms FHProfUnd an der Hochschule Niederrhein unter Leitung von Prof. Dr. Hirsch, Förderkennzeichen 1762X08

Zusammenfassung

Im 4. Projektabschnitt wurden die Erkennungsergebnisse der im 2. Projektbericht definierten Experimente analysiert und verglichen. Diese wurden anhand der bereits beschriebenen Erkennungssysteme, die mit einer Adaption der Referenzmuster und einer robusten Merkmalsextraktion arbeiten, erstellt. Weiterhin wurden Experimente mit „multi-mode“-Modellen durchgeführt. Aus der Analyse und dem Vergleich der Ergebnisse soll abgeleitet werden, welche Möglichkeiten zur Kombination der Erkennungssysteme bestehen. Zur Ideenfindung von möglichen Ansätzen zur Kombination von verschiedenen Erkennungsfrontends wurde eine Literaturrecherche betrieben, deren Ergebnis im Anhang des vorliegenden Textes zusammengefasst ist. Darin wird ein Überblick über verschiedene Möglichkeiten zur Kombination gegeben, die in wissenschaftlichen Veröffentlichungen auf Fachtagungen und in verschiedenen Fachzeitschriften beschrieben werden.

Untersuchung und Analyse der Erkennungsergebnisse

Zu den bereits im 2. und 3. Projektbericht[1] vorgestellten Erkennungsergebnisse, die unter Verwendung der beiden robusten Erkennungssysteme (System 1, Robuster Erkenner, Adaption der Referenzmuster, kurz „HGH adapt“ und System 2, Robuster Erkenner, robuste Merkmalsextraktion, kurz „HGH robust“) arbeiten, sollen im Folgenden weitere Untersuchungen mit so genannten „multi-mode“- Modellen durchgeführt werden, deren Ergebnisse mit denen der robusten Erkennungssysteme verglichen werden. Als Grundlage für die neuen und bereits durchgeführten Untersuchungen dienen die bereits definierten Experimente, die auf der Basis der Sprachdaten der „Aurora 5“ Datenbank aufbauen. Diese sind nachfolgend noch einmal dargestellt:

- Testcase „Clean“ beinhaltet ungestörte Daten.
- Testcase „G712 CarNoise“ beinhaltet Daten, die mittels Hintergrundgeräuschen aus Fahrzeugen in versch. SNR-Abstufungen (0dB, 5dB, 10dB und 15dB) gestört wurden. Zusätzlich wurden die Daten gefiltert, um einen Telefonkanal zu simulieren (G712- Filtercharakteristik).
- Testcase „G712 CarNoise Handsfree“ beinhaltet die gleichen Daten wie Testcase „G712 CarNoise“. Zusätzlich wurden die Rohdaten mit einer Raumimpulsantwort gefaltet um Freisprechen im Fahrzeug zu simulieren.
- Testcase „IntNoise“ beinhaltet Daten, die mittels Hintergrundgeräuschen aus Büros, Lobbies etc. in versch. SNR-Abstufungen (5dB, 10dB und 15dB) gestört wurden.
- Testcase „G712 CarNoise Handsfree“ beinhaltet die gleichen Daten wie Testcase „IntNoise“. Zusätzlich wurden die Rohdaten mit einer Raumimpulsantwort gefaltet um Freisprechen in Räumen zu simulieren.

Für die Erkennungsexperimente, die mittels der „multi-mode“ Referenzmustern durchgeführt werden, wird dieselbe Cepstralanalyse verwendet, die auch bei den „HGH adapt“-Experimenten zum Einsatz kommt, jedoch ohne Verwendung der Adaption. Die „multi-mode“- Experimente werden kurz als „HGH multi-mode“ bezeichnet, die dazugehörigen Referenzen wurden anhand von ungestörten und zuvor gestörten Trainingsdaten erzeugt. Der „multi-mode“- Ansatz soll verhindern, dass es bei der Erkennung von gestörten Sprachsignalen durch die Abweichung der Trainingsdaten von den Testdaten zu einer Verschlechterung der Erkennungsrate kommt. Dazu werden für das Training der Modelle Sprachdaten aus unterschiedlichen Störumgebungen verwendet, die auch bei gestörten Testdaten eine gute Anpassung der Modelle an die Störsituation gewährleisten.

[1] H.G. Hirsch, Arbeitsberichte zum 2. und 3. Projektabschnitt, verfügbar unter http://dnt.kr.hsnr.de

In Tabelle 1 sind die verschiedenen Umgebungen, die in das Training mit einfließen, dargestellt:

Set A	Set B
G712_HandsfreeCar	InteriorNoise00_Handsfree Office / 05 / 10 / 15
G712_GSM	InteriorNoise00_Handsfree Livingroom / 05 / 10 / 15
G712_StreetNoise00_GSM / 05 / 10 / 15	InteriorNoise00 / 05 / 10 / 15
G712_HandsfreeCar_GSM	HandsfreeOffice
G712_CarNoise00_Handsfree Car_GSM / 05 / 10 / 15	HandsfreeLivingroom
G712_CarNoise00_Handsfree Car / 05 / 10 / 15	Clean
G712_CarNoise00 / 05 / 10 / 15	—
G712	—

Tabelle 1: Störumgebungen für „multi-mode“- Referenzen

Die Anzahl der Dateien, die für das Training der Referenzmuster Set A und Set B verwendet wurden, entspricht der Anzahl der Dateien des Trainings für die „Clean“-Referenzmuster (8623 Dateien). Die Aufspaltung in Set A und Set B soll zum einen eine bestmögliche Erkennung in den beiden Haupt- Testumgebungen „G712 CarNoise“ und „InteriorNoise“ garantieren. Zum anderen wird damit der Überlegung Sorge getragen, dass Experimente durchgeführt werden können, bei denen die kritische Situation auftritt, dass bei der Erkennung eine Spracheingabe in einer akustischen Umgebung verwendet wird, die nicht im „multi-mode“ Training berücksichtigt wurde. Set A wurde für die Erkennung von gestörten Daten aus dem KFZ Bereich inkl. Telefonkanal- Simulation (G712 CarNoise - Segment) erzeugt. Set B dient der Erkennung von Daten, die Umgebungs- und/oder Alltagsgeräusche enthalten (InteriorNoise gestörte Daten). In beiden Sets wird neben Nahbesprechung auch Freisprechen berücksichtigt.
In den Trainingsdaten wurden neben den in den Experimenten definierten Störumgebungen auch Daten verwendet, die nicht direkt in den Experimenten auftreten, jedoch der Störumgebung in den Experimenten ähneln. Für das Training von Set B wurden neben den in den Experimenten definierten SNR-Stufen von 5 bis 15dB noch zusätzlich Trainingsdaten mit einem SNR von 0dB verwendet.
Die Ergebnisse der verschiedenen Experimente sind in Form von Wortfehlerraten in den Tabellen 2 bis 6 dargestellt.

Bei der Betrachtung der Ergebnisse wird deutlich, dass die Ergebnisse der Erkennungsexperimente der beiden robusten Erkenner vergleichbare Wortfehlerraten liefern. In einzelnen Fällen überwiegt das eine System das Andere etwas stärker, jedoch ist im Mittel die Wortfehlerrate vergleichbar. Eine genaue Betrachtung der einzelnen Ergebnisse befindet sich im 3. Projektbericht.
Zusätzlich zu den bereits betrachteten Ergebnissen kommen noch die Ergebnisse der Erkennung mit „multi-mode“- Modellen hinzu. Diese sollen im Folgenden eingehender betrachtet werden, um einen Vergleich zwischen der Erkennung mit den robusten Verfahren und der Erkennung mit „multi-mode“- Modellen zu ermöglichen.

Erkenner	Clean
HGH adapt	0,53 %
HGH robust	0,54 %
HGH multi-mode Set A	0,79 %
HGH multi-mode Set B	1,10 %

Tabelle 2: Wortfehlerraten Clean

Aus der Betrachtung der Ergebnisse für den Testfall „clean“ (als Testdaten wurden ungestörte Daten verwendet) in Tabelle 2 wird deutlich, dass die Verwendung von beiden „multi-mode“- Modellsätzen in diesem Testfall zu einer Erhöhung der Wortfehlerrate im Vergleich zu den beiden robusten Systemen führt. Dies ist auf die Abweichung der ungestörten Testdaten von den gestörten Trainingsdaten, aus denen die Modelle erzeugt wurden, zurückzuführen. Weiterhin ist ersichtlich, dass sich der Anteil der gestörten Daten in Set B stärker ausprägt als in Set A, da bei der Erkennung mit Set B die Wortfehlerrate deutlich höher liegt als bei der Erkennung mit Set A.
In Tabelle 3 sind die Ergebnisse der Experimente für die „G712 CarNoise“ Testdaten dargestellt:

Erkenner	G712Car-Noise 0dB	G712Car-Noise 5dB	G712Car-Noise 10dB	G712Car-Noise 15dB
HGH adapt	18,73 %	5,83 %	2,11 %	1,15 %
HGH robust	17,45 %	5,87 %	2,24 %	1,23 %
HGH multi-mode Set A	11,49 %	4,02 %	2,12 %	1,52 %
HGH multi-mode Set B	12,69 %	4,96 %	2,63 %	1,85 %

Tabelle 3: Wortfehlerraten G712 CarNoise

Aus den Wortfehlerraten dieser Experimente wird deutlich, dass im SNR Bereich von 0dB bis 5dB die Anwendung der „multi-mode“- Modelle zu einer deutlichen Senkung der

Wortfehlerrate im Vergleich zu den robusten Systemen führt. Selbst die Verwendung von Set B, das nicht auf die Störumgebung in den Testdaten hin trainiert wurde, führt zu einer, im Vergleich mit den beiden robusten Verfahren, gesenkten Wortfehlerrate. Dies ist im Bezug zu „HGH adapt“ auf eine noch bessere Anpassung der Modelle an die Störumgebung zurückzuführen als es mit einer Adaption der Referenzmodelle möglich ist. Im Bezug auf „HGH robust“ ist zu vermuten, dass es, bedingt durch einen Restanteil von Störungen im Testsignal nach der robusten Merkmalsextraktion und der Verwendung von ungestörten „Clean“- Modellen für die Erkennung und der damit verbundenen leichten Fehlanpassung zwischen Modellen und Testdaten zu keiner weiteren Senkung der Wortfehlerrate kommt. Im oberen SNR- Bereich von 10dB und 15dB liefert „HGH adapt“ die geringste Wortfehlerrate, wobei die anderen Verfahren keine großen Differenzen zur Wortfehlerrate von „HGH adapt“ aufweisen.
Tabelle 4 enthält die Ergebnisse der Erkennungsexperimente für die „G712 CarNoise Handsfree“ Testdaten:

Erkenner	G712Car-Noise Handsfree 0dB	G712Car-Noise Handsfree 5dB	G712Car-Noise Handsfree 10dB	G712Car-Noise Handsfree 15dB
HGH adapt	41,74 %	15,26 %	4,93 %	2,09 %
HGH robust	35,66 %	14,57 %	5,99 %	3,00 %
HGH multi-mode Set A	27,87 %	9,42 %	3,78 %	2,16 %
HGH multi-mode Set B	35,90 %	16,80 %	8,54 %	5,44 %

Tabelle 4: Wortfehlerraten G712 CarNoise Handsfree

Auch bei diesen Experimenten kommt es durch die Verwendung der „multi-mode“- Modelle aus Set A in einem SNR- Bereich von 0dB bis 10dB zu einer teils erheblichen Senkung der Wortfehlerrate im Vergleich zu den beiden robusten Erkennungssystemen. Diese Senkung ist auf die bereits weiter oben genannten Gründe zurückzuführen. Dies wird gerade im Bezug auf den Nachhall deutlich, da dieser durch die Verwendung der „multi-mode“- Modelle besser berücksichtigt werden kann als z.B. bei der Adaption der Referenzmuster bei „HGH adapt“.
Das Störgeräusch-fremde Set B hingegen führt nur im unteren SNR- Bereich von 0dB zu einer besseren Wortfehlerrate bezogen auf „HGH adapt“, in allen anderen Fällen und im Bezug auf „HGH robust“ ist die Wortfehlerrate schlechter. Bei einem SNR von 15dB kann durch den Einsatz von „HGH adapt“ die geringste Wortfehlerrate erreicht werden.
In Tabelle 5 sind die Ergebnisse der Experimente mit „InteriorNoise“- Testdaten dargestellt:

Erkenner	IntNoise 5dB	IntNoise 10dB	IntNoise 15dB
HGH adapt	13,81 %	4,89 %	2,14 %
HGH robust	15,75 %	6,36 %	2,92 %
HGH multi-mode Set A	18,39 %	9,56 %	5,75 %
HGH multi-mode Set B	16,11 %	8,64 %	4,80 %

Tabelle 5: Wortfehlerraten InteriorNoise

Hierbei ist auffällig, dass in diesem Experiment zwar das der Störsituation entsprechende „multi-mode“- Set B geringere Wortfehlerraten liefert als das störgeräuschfremde Set A, es wird aber auch deutlich, dass beide „multi-mode“- Sets über den gesamten SNR-Bereich höhere Wortfehlerraten als die robusten Erkennungssysteme aufweisen. Dies kann an der Verwendung eines hohen Anteils an verhallten gestörten Trainingsdaten bei der Erzeugung der Modelle für Set B liegen, die bei diesen Experimenten zu einer zu großen Abweichung zwischen den Modellen zur Erkennung und den Testdaten führen.
Die Wortfehlerraten für eine Erkennung mit „InteriorNoise Handsfree“- Testdaten sind in Tabelle 6 zusammengestellt:

Erkenner	IntNoise 5dB	IntNoise 10dB	IntNoise 15dB
HGH adapt	27,15 %	12,46 %	6,38 %
HGH robust	27,64 %	13,90 %	7,83 %
HGH multi-mode Set A	27,90 %	14,49 %	8,57 %
HGH multi-mode Set B	19,11 %	9,95 %	5,53 %

Tabelle 6: Wortfehlerraten InteriorNoise Handsfree

Bei der Betrachtung dieser Ergebnisse wird deutlich, dass hier das der Störsituation entsprechende „multi-mode“- Set B im Vergleich zu den anderen Systemen über den gesamten SNR-Bereich geringere Wortfehlerraten liefert. Dies ist nach der Betrachtung der Ergebnisse für das „multi-mode“ Set B in Tabelle 5 zu erwarten, da, wie bereits oben schon angemerkt, ein großer Anteil der Trainingsdaten zu Set B aus dem „Handsfree“- Bereich stammt, was ein hohes Maß an Anpassung zwischen den Modellen und den halligen Testdaten ermöglicht. Das Störgeräusch-fremde Set A liefert im Vergleich zu den robusten Systemen eine höhere Wortfehlerrate, was erneut auf eine mangelnde Anpassung bedingt durch die Verwendung der nicht angepassten Störgeräusche und verhallten Daten aus dem KFZ- Bereich im Modell- Set A zurückzuführen ist.

Eine weitere Möglichkeit zur Senkung der Wortfehlerrate ist die kombinierte Verwendung von robusten Merkmalen, die mittels „HGH robust“ erstellt wurden und „multi-

mode“ Modellen. Dieser Ansatz wird im folgenden Abschnitt untersucht.
Für das Training der „multi-mode“- Modelle wurden die selben Trainingsdaten (siehe Tabelle 1) wie bereits bei den „HGH multi-mode“ Experimenten verwendet, um eine Vergleichbarkeit zu gewährleisten. Die Ergebnisse für die einzelnen Experimente sind in den Tabellen 7 bis 11 aufgeführt. Um einen besseren Vergleich zu ermöglichen, sind die Ergebnisse aus den Tabellen 2 bis 6 noch einmal mit aufgeführt. In Tabelle 7 sind die Erkennungsergebnisse für das Experiment mit ungestörten Daten aufgelistet:

Erkenner	Clean
HGH adapt	0,53 %
HGH robust	0,54 %
HGH multi-mode Set A	0,79 %
HGH multi-mode Set B	1,10 %
HGH robust multi-mode Set A	0,97 %
HGH robust multi-mode Set B	0,96 %

Tabelle 7: Wortfehlerraten Clean

Wie bereits bei den Ergebnissen der Erkennung mit den einfachen „HGH multi-mode“-Modellen ist auch bei dem kombinierten Ansatz „HGH robust multi-mode“ keine Senkung der Wortfehlerrate im Vergleich zu den Ergebnissen, die die robusten Erkennungssysteme mit ungestörten Modellen liefern, möglich. Dies ist, wie weiter oben bereits angemerkt wurde, auf die Abweichung der ungestörten Testdaten von den gestörten Trainingsdaten, aus denen die Modelle erzeugt wurden, zurückzuführen.

Tabelle 8 beinhaltet die Ergebnisse für die „G712 CarNoise“- Testdaten:

Erkenner	G712Car-Noise 0dB	G712Car-Noise 5dB	G712Car-Noise 10dB	G712Car-Noise 15dB
HGH adapt	18,73 %	5,83 %	2,11 %	1,15 %
HGH robust	17,45 %	5,87 %	2,24 %	1,23 %
HGH multi-mode Set A	11,49 %	4,02 %	2,12 %	1,52 %
HGH multi-mode Set B	12,69 %	4,96 %	2,63 %	1,85 %
HGH robust multi-mode Set A	9,30 %	3,24 %	1,62 %	1,13 %
HGH robust multi-mode Set B	9,80 %	4,05 %	2,08 %	1,46 %

Tabelle 8: Wortfehlerraten G712 CarNoise

Aus den Ergebnissen geht hervor, dass in diesem Block von Experimenten die Anwendung des kombinierten Ansatz „HGH robust multi-mode“ mit dem Modell-Set A zu einer weiteren Reduktion der Wortfehlerrate im Vergleich zu den einfachen „multi-Mode“- Modellen und den robusten Systemen mit „clean“- Modellen über den gesamten SNR- Bereich

führt. Dies ist darauf zurückzuführen, dass der durch die Robuste Merkmalsextraktion verbleibende Störanteil in den Testsignalen und die aus gestörten Daten aus ähnlichen Störszenarien trainierten Modellen ein hohes Maß an Anpassung aufweisen, was zu einer Verbesserung der Erkennung und damit zu einer Senkung der Wortfehlerrate führt. Dieser Effekt ist auch bei dem Set B der „HGH robust multi-mode“ zu beobachten. Im Vergleich „HGH multi-mode“ und „HGH robust multi-mode“ wird deutlich, dass noch eine weitere Reduktion der Wortfehlerrate durch Anwendung des kombinierten Ansatzes möglich ist.

In Tabelle 9 sind die Ergebnisse für die „G712 CarNoise Handsfree“- Experimente dargestellt:

Erkenner	G712Car-Noise Handsfree 0dB	G712Car-Noise Handsfree 5dB	G712Car-Noise Handsfree 10dB	G712Car-Noise Handsfree 15dB
HGH adapt	41,74 %	15,26 %	4,93 %	2,09 %
HGH robust	35,66 %	14,57 %	5,99 %	3,00 %
HGH multi-mode Set A	27,87 %	9,42 %	3,78 %	2,16 %
HGH multi-mode Set B	35,90 %	16,80 %	8,54 %	5,44 %
HGH robust multi-mode Set A	20,99 %	6,95 %	2,99 %	1,74 %
HGH robust multi-mode Set B	23,54 %	9,66 %	4,95 %	2,91 %

Tabelle 9: Wortfehlerraten G712 CarNoise Handsfree

Auch bei diesen Experimenten kommt es durch den Einsatz des „HGH robust multi-mode“ Modell-Set A zu einer erheblichen Reduktion der Wortfehlerrate über den gesamten SNR-Bereich im Vergleich zu den anderen Experimenten, die Verwendung des Störgeräusch-fremden Set B führt im Vergleich der Ergebnisse der robusten Systeme und der Ergebnisse der Erkennung mit „HGH multi-mode“- Set B auch zu einer Senkung der Wortfehlerrate.

Tabelle 10 enthält die Ergebnisse der Experimente mit „IntNoise“- gestörten Daten:

Erkenner	IntNoise 5dB	IntNoise 10dB	IntNoise 15dB
HGH adapt	13,81 %	4,89 %	2,14 %
HGH robust	15,75 %	6,36 %	2,92 %
HGH multi-mode Set A	18,39 %	9,56 %	5,75 %
HGH multi-mode Set B	16,11 %	8,64 %	4,80 %
HGH robust multi-mode Set A	13,95 %	6,99 %	3,82 %
HGH robust multi-mode Set B	14,21 %	6,92 %	3,55 %

Tabelle 10: Wortfehlerraten InteriorNoise

Bei der kombinierten Anwendung der robusten Merkmalsextraktion „HGH robust“ und den „multi-mode“- Modellen kommt es bei diesem Block von Experimenten nicht zu einer weiteren Senkung der Wortfehlerrate im Vergleich zum robusten System „HGH adapt“. Lediglich die Wortfehlerrate von „HGH robust“ ohne „multi-mode“- Modelle konnte im unteren SNR- Bereich von 5dB gesenkt werden. Diese Tatsache ist, wie bereits weiter oben schon geäußert wurde, auf die sehr halllastige Zusammensetzung der gestörten Daten für das Training zurückzuführen, die bei diesem Experiment zu einem geringen Grad der Anpassung zwischen den „multi-mode“- Modellen und den Testdaten führt. Dies wird auch an den Ergebnissen aus Set A für den gesamten SNR-Bereich deutlich. Hier resultiert bei der Verwendung des störumgebungfernen Modell-Sets eine höhere Wortfehlerrate im Vergleich zu „HGH adapt“.

In Tabelle 11 sind die Ergebnisse der Experimente mit „IntNoise Handsfree“- Testdaten zusammengefasst:

Erkenner	IntNoise Handsfree 5dB	IntNoise Handsfree 10dB	IntNoise Handsfree 15dB
HGH adapt	27,15 %	12,46 %	6,38 %
HGH robust	27,64 %	13,90 %	7,83 %
HGH multi-mode Set A	27,90 %	14,49 %	8,57 %
HGH multi-mode Set B	19,11 %	9,95 %	5,53 %
HGH robust multi-mode Set A	19,30 %	9,77 %	5,67 %
HGH robust multi-mode Set B	16,00 %	7,67 %	4,31 %

Tabelle 11: Wortfehlerraten InteriorNoise Handsfree

Bei diesen Experimenten ist wieder eine Senkung der Wortfehlerrate bei kombinierter Verwendung von „HGH robust“ und den „multi-mode“- Modellen des störszenarionahen Set B zu beobachten. Auch die Verwendung des störszenariofernen Set A bringt eine Senkung der Wortfehlerrate im Vergleich zu der Anwendung der beiden robusten Systeme und den Ergebnissen der Erkennung mit „HGH multi-mode“ -Set A. Die weitere Reduktion der Wortfehlerrate bei Verwendung der Kombination aus „HGH robust“ und den „multi-mode“- Modellen Set B ist auf den hohen Anteil an verhallten Trainingsdaten und ein daraus folgender hoher Grad an Anpassung zwischen den „multi-mode“- Modellen und den aus den Testdaten extrahierten robusten Merkmalen zurückzuführen.

Fazit „Auswertung der Erkennungsergebnisse“

Als Fazit aus der Betrachtung der Erkennungsergebnisse kann ganz klar gefolgert werden, dass eine Anwendung eines einfachen „multi-mode“ Systems oder die Kombination

„HGH robust“ mit „multi-mode“- Modellen eine zusätzliche Senkung der Wortfehlerrate im Vergleich zu der Anwendung der beiden robusten Systeme mit ungestörten „Clean“-Modellen zur Folge hat. Jedoch dürfen hierbei der mit den „multi-mode“- Modellen verbundene Aufwand bei der Erstellung der Modelle und weitere Restriktionen nicht außer Acht gelassen werden. Zur sinnvollen Erzeugung der „multi-mode“- Modelle sind detaillierte Informationen über die störbehafteten Umgebungen, in der der Erkenner eingesetzt werden soll, notwendig.
Das bedeutet, es müssen im Idealfall Trainingsdaten, die in den entsprechenden Einsatzumgebungen aufgenommen wurden, vorliegen. Besteht diese Möglichkeit nicht, sollten zumindest ausreichend viele Proben mit den entsprechenden Störgeräuschen aus den Umgebungen, in denen das System später eingesetzt werden soll, sowie Messungen zur Bestimmung der akustischen Umgebung vorliegen, um anhand dieser Daten gestörte Trainingsdaten zu erzeugen.
Ein zusätzliches Problem, dass bei der Verwendung von „multi-mode“- Modellen auftritt, ist die Restriktion der Einsatzumgebung. Aus den Ergebnissen in den Tabellen 2 bis 6 wird deutlich, dass es zwar durchaus möglich ist, ein Modell- Set, das nicht für die passende Störumgebung trainiert wurde, erfolgreich einzusetzen. Jedoch wird z.B. aus den Ergebnissen aus Tabelle 5 deutlich, dass sowohl die störgeräuschnahen als auch die störgeräuschfernen Modelle zu keiner Senkung der Wortfehlerrate führen. Die beiden robusten Systeme hingegen leisten auch in dieser Testumgebung eine zuverlässige Erkennung, bedingt durch ihre universelle Einsetzbarkeit.
Zusammenfassend wird deutlich, dass die Anwendung der „multi-mode“- Modelle im Vergleich zu den beiden robusten Systemen die besten Erkennungsraten liefert. Es empfiehlt sich jedoch aus den bereits oben aufgeführten Gründen nur bedingt, diese Modelle einzusetzen.
Als sinnvoll gestaltet sich die Anwendung der „multi-mode“- Modelle, wenn das Störszenario bereits vor der Erkennung feststeht. Das bedeutet, es tritt eine begrenzte Anzahl an bereits bekannten Störsituationen auf zu der jeweils passende „multi-mode“- Modelle trainiert werden können. Die entsprechenden Modellsätze können dann vor der eigentlichen Erkennung durch einen Funktionsblock zur Selektion ausgewählt und anschließend zur Erkennung verwendet werden. Untersuchungen zu diesem Ansatz wurden im Rahmen des Projektes in der Masterthesis „Robuste Erkennung gestörter Sprachsignale mit in gestörter Umgebung trainierten Referenzmustern und einer Adaption auf die individuellen Störbedingungen“[2] durchgeführt.
Im weiteren Verlauf wird der „multi-mode“- Ansatz durch seine eingeschränkte Verwend-

[2] A. Kitzig, „Robuste Erkennung gestörter Sprachsignale mit in gestörter Umgebung trainierten Referenzmustern und einer Adaption auf die individuellen Störbedingungen“, Masterthesis, verfügbar unter http://dnt.kr.hsnr.de

barkeit nicht näher untersucht und der Fokus wird, bedingt durch ihre universelle Anwendbarkeit, auf die beiden robusten Erkennungssysteme gelegt.

Detaillierte Betrachtung der Erkennungsergebnisse

Nach der Betrachtung der Wortfehlerraten für die einzelnen Experimente soll im folgenden Teil näher untersucht werden, wie sich die einzelnen Erkennungsergebnisse der beiden robusten Erkennungssysteme im Detail unterscheiden. Daraus soll eine Grundlage für weitere Untersuchungen im Bereich der Kombination von verschiedenen Erkennungssystemen geschaffen werden.

Es werden alle im 2. Projektbericht aufgeführten Experimente für die „Aurora 5“ Datenbank (8700 Testdateien je Störumgebung) für „HGH adapt“ und „HGH robust“ unter Verwendung entsprechender „Clean“- Modelle auf Basis der einzelnen Erkennungsergebnisse untersucht. Dabei werden folgende Unterteilungen bei der Betrachtung der Ergebnisse vorgenommen:

- beide Erkenner liefern ein gleiches, korrektes Erkennungsergebnis (Fall 1)
- ein Erkenner liefert ein korrektes Ergebnis, der andere Erkenner ein falsches Ergebnis (Fall 2)
- beide Erkenner liefern fehlerhafte Ergebnisse
  - beide Erkenner liefern den gleichen Fehler (Fall 3)
  - die Fehler der Erkenner unterscheiden sich (Fall 4)

Besonders interessant für eine mögliche Kombination der beiden Systeme ist der Fall 2, bei dem ein System ein richtiges, das andere System ein falsches Ergebnis liefert. Weiterhin lässt sich in den Fällen 3 und 4 zeigen, dass sich die beiden Systeme auch bei fehlerhaften Ergebnissen unterscheiden. Die Betrachtung von Fall 3 ist dabei nur der Vollständigkeit halber aufgelistet, da die Auswertung von Fall 3 keine weitere Aussage ermöglicht. Bei Fall 4 liegen, ähnlich dem Fall 2, zunächst zwei unterschiedliche Erkennungsergebnisse vor, die jedoch beide fehlerhaft sind. In Fall 4 besteht daher keine direkte Möglichkeit, ein korrektes Erkennungsergebnis nur durch die Kombination der beiden Systeme zu erhalten, jedoch könnte bei diesem Fall mittels einer erweiterten Signalverarbeitung oder einem modifizierten Erkennungssystem versucht werden, durch gezielte Nachverarbeitung ein verbessertes Erkennungsergebnis zu erzielen.
Vor der Darstellung der Auswertung über die Zusammensetzung der Erkennungsergebnisse wird vorab noch kurz erläutert, wie sich der Ansatz zum Vergleich der Ergebnisse aufbaut. Dazu wird zuerst der Aufbau der Datei, die die Erkennungsergebnisse enthält,

betrachtet. Die Ergebnisse der Experimente liegen nach der Erkennung als Textdateien mit der Endung „.rec“ vor. Als Beispiel sind die Erkennungsergebnisse für die Äußerung „zero five nine“ eines Sprechers (z59.rah.raw) in der störbehafteten Testumgebung „G712_CarNoise05“ für beide robusten Erkennungssysteme dargestellt:

```
Erkennungsergebnis HGH adapt:
        0   2875000 w_sil
  2875000   5975000 w_z
  5975000   6475000 w_sil
  6475000   8875000 w_5
  8875000   9375000 w_sil
  9375000  12275000 w_9
 12275000  15075000 w_sil

Erkennungsergebnis HGH robust:
        0   3875000 w_sil
  3875000   6475000 w_o
  6475000   9575000 w_5
  9575000  12275000 w_9
 12275000  15075000 w_sil
```

Die Dateien mit den Ergebnissen enthalten neben den erkannten Wörtern inklusive erkannter Pausen (Modell w_sil) auch noch die dazugehörigen zeitlichen Informationen. Deutlich wird bei der Betrachtung der beiden Ergebnisse im Beispiel, dass die Erkennung mit „HGH adapt“ zu einem korrekten Ergebnis führt, „HGH robust“ liefert bei der Erkennung der ersten Ziffer ein fehlerhaftes Ergebnis.
Anhand dieser Ergebnisse wird nun ein Vergleich der Ergebnisse der beiden Erkennungssysteme durchgeführt. Dabei werden die Ergebnisse nicht nach einzelnen Ziffern sondern auf Satz-Ebene betrachtet, d.h., vollständig korrekt erkannte Ergebnisse werden als richtig, vollständig oder teilweise falsche erkannte Ergebnisse werden als falsch betrachtet, wobei die erkannten Pausen nicht in die Auswertung mit einfließen.
Die Auswertung der Ergebnisse der beiden Systeme wurde unter Verwendung des Moduls „HResults“ aus dem HTK- Toolkit[3] durchgeführt, welches in einem Matlab- Code zur Durchführung des Vergleichs eingebunden wurde.
Die Ergebnisse des Vergleiches sind zur besseren Übersicht in mehrere Tabellen aufgeteilt und in den Tabellen 12 bis 14 dargestellt. Dabei beschreibt die jeweilige Zahl in den Spalten

[3]S. Young, The HTK book, Cambridge University, 2006

der Tabelle die Anzahl der in dem entsprechenden Testfall enthaltenen Sprachfiles.

Experiment	beide fehlerhaft		HGH adapt korrekt / HGH robust fehlerhaft	HGH adapt fehlerhaft /HGH robust korrekt
	gleicher Fehler	unterschiedlicher Fehler		
Clean	97	10	35	27

Tabelle 12: Ergebnisvergleich „clean“

In Tabelle 12 sind die Ergebnisse des Vergleichs von Erkennungen mit „Clean“- Daten dargestellt. Aus der Betrachtung der Ergebnisse wird deutlich, dass sich neben dem Fall, dass beide Erkenner ein fehlerhaftes Ergebnis liefern (Fall 3 und 4, insgesamt 107 Dateien) auch noch ein gewisses Potential für eine Unterscheidung in fehlerhafte/fehlerfreie Erkennung der beiden Systeme ergibt (Fall 2). So liefert „HGH adapt“ in 35 Fällen ein korrektes Ergebnis, während „HGH robust“ bei den selben Sprachdaten ein fehlerhaftes Ergebnis liefert und in 27 Fällen liefert „HGH robust“ ein korrektes Ergebnis während „HGH adapt“ einen Fehler bei der Erkennung aufweist. Somit liegt selbst in dieser Testumgebung, die eine sehr geringe Wortfehlerrate aufweist („HGH adapt“ => 0,53 % und „HGH robust“ => 0,54 %) in 62 weiteren Fällen eine Möglichkeit, die Wortfehlerrate durch eine Kombination der Erkenner zu senken.
Tabelle 13 enthält die Ergebnisse des Vergleiches für die „G712 CarNoise“- Testdaten:

Experiment	beide fehlerhaft		HGH adapt korrekt / HGH robust fehlerhaft	HGH adapt fehlerhaft /HGH robust korrekt
	gleicher Fehler	unterschiedlicher Fehler		
G712 CarNoise 0dB	720	1732	973	997
G712 CarNoise 0dB Handsfree	835	3831	750	965
G712 CarNoise 5dB	376	379	631	582
G712 CarNoise 5dB Handsfree	626	1365	941	973
G712 CarNoise 10dB	183	100	272	239
G712 CarNoise 10dB Handsfree	300	388	693	467
G712 CarNoise 15dB	121	39	146	138
G712 CarNoise 15dB Handsfree	180	138	407	189

Tabelle 13: „G712 CarNoise“

Bei diesen Daten ist neben den Fällen 3 und 4, bei denen beide Erkenner fehlerhafte Ergebnisse liefern, eine klare Unterscheidung in eine fehlerhafte/fehlerfreie Erkennung (Fall 2) der beiden Erkennungssysteme möglich. Bei den gestörten Testdaten ist die Anzahl der fehlerhaften/fehlerfreien Erkennung (Fall 2) der beiden Systeme bedingt durch die höhere Wortfehlerrate im Vergleich zu den Experimenten mit den ungestörten „clean“- Testdaten noch ausgeprägter. So würde zum Beispiel im Idealfall eine erfolgreiche Kombination der beiden Systeme im Testfall „G712 CarNoise 0dB“ eine Anzahl von 1970 weiteren Sprachäußerungen korrekt erkannt werden, was eine erhebliche Senkung der Wortfehlerrate zur Folge hätte. Ein ähnlich ausgeprägtes Ergebnis ist bei den Vergleichen der „InteriorNoise“-Testdaten in Tabelle 14 zu sehen:

Experiment	beide fehlerhaft		HGH adapt korrekt / HGH robust fehlerhaft	HGH adapt fehlerhaft /HGH robust korrekt
	gleicher Fehler	unterschiedlicher Fehler		
Interior Noise 5dB	383	1513	1347	950
Interior Noise 5dB Handsfree	346	3085	1269	1146
Interior Noise 10dB	272	396	834	495
Interior Noise 10dB Handsfree	319	1289	1223	963
Interior Noise 15dB	174	130	422	230
Interior Noise 15dB Handsfree	305	571	885	602

Tabelle 14: „InteriorNoise“

Auch in diesen Testumgebungen kommt es zu einer deutlichen Unterscheidung der Fälle „beide fehlerhaft“ (Fall 3 und 4) und „ein System fehlerfrei, das andere System fehlerhaft“. Somit liegt auch in dieser Testumgebung ein hohes Kombinations- Potential vor.

Fazit „Vergleich der Erkennungsergebnisse“

Als Fazit aus dem Vergleich der Erkennungsergebnisse kann gefolgert werden, dass ein hohes Potential für die Kombination der beiden Erkennungssysteme vorliegt, bedingt durch die hohe Anzahl an fehlerhafter/fehlerfreier Erkennung (Fall 2) für die jeweiligen Systeme „HGH adapt“ und „HGH robust“ in jeder Testumgebung. Im vorangegangenen Text wurde immer von einem „Idealfall“, d.h., die Kombination liefert ein einhundert Prozent korrektes Ergebnis, ausgegangen, was nicht unbedingt realistisch ist, jedoch zeigen die Ergebnisse, das bei einer erfolgreichen Kombination der beiden Systeme eine deutliche Senkung der Wortfehlerrate resultieren könnte.

Die vorliegende Untersuchung stützt sich auf die Auswertung der Ergebnisse auf Satzebene und vermittelt einen guten Überblick über das vorhandene Kombinations- Potential. Eine weitere Möglichkeit der Betrachtung ist die Auswertung der Erkennungsergebnisse im Bereich der Wortebene. Da sich aber bereits bei der Auswertung der Ergebnisse auf Satz- Ebene ein großes Kombinations- Potential herausgestellt hat, wird dieser Ansatz an

dieser Stelle nicht weiter verfolgt.

Im nächsten Projektabschnitt soll nun ein geeigneter Ansatz zur Kombination der beiden robusten Erkennungssysteme „HGH adapt“ und „HGH robust“ gefunden werden. Ein Überblick über mögliche Ansätze zur Kombination, die bereits von anderen Autoren erfolgreich umgesetzt werden konnten, ist im Anschluss im Anhang dieses Textes gegeben.

Anhang

Aus dem vorherigen Teil des Berichts wurde deutlich, dass die Verwendung von verschiedenen Erkennungssystemen zur Robusten Erkennung bei gleichen Testumgebungen ähnliche Erkennungsraten liefert, es konnte jedoch auch festgestellt werden, dass sich, bedingt durch die unterschiedlichen Vorgehensweisen, Fehler bei der Erkennung unterschiedlich verteilen. Diese Tatsache ermöglicht einen zusätzlichen Ansatzpunkt zur Reduktion der Wortfehlerrate, in dem die verschiedenen Erkennungssysteme miteinander kombiniert werden. In verschiedenen wissenschaftlichen Veröffentlichungen konnte gezeigt werden, das durch die Kombination von unterschiedlichen Erkennungssystemen eine Reduktion der Wortfehlerrate möglich ist. Um einen Überblick über die verschiedenen Ansatzmöglichkeiten zur Kombination zu erhalten, wurde eine Literaturrecherche durchgeführt, die folgende Veröffentlichungen aus Fachzeitschriften und Fachtagungen beinhaltet:

[AK03] Waleed H. Abdulla, Nikola Kasabov. Reduced feature-set based parallel chmm speech recognition systems. *Information Sciences*, 156:21–38, 2003.

[Bur04] Lukáš Burget. Combination of speech features using smoothed heteroscedastic linear discriminant analysis. *ICSLP 2004*, 2004.

[EB00] Daniel P. W. Ellis, Jeff A. Bilmers. Using mutual information to design feature combinations. *Int. Conf. on Spoken Language Processing*, strony 79–82, 2000.

[Ell00a] Daniel P. W. Ellis. Improved recognition by combining different features and different systems. *Proc. of the AVIOS-2000*, 2000.

[Ell00b] Daniel P. W. Ellis. Stream combination before and/or after acoustic model. *ICASSP-2000*, 2000.

[ER01] Daniel P. W. Ellis, Manuel J. Reyes Gomez. Investigations into tandem acoustic modelling for aurora task. *Proceedings Eurospeech*, 2001.

[Fis97] Jonathan G. Fiscus. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (rover). *IEEE Workshop on Automatic Speech Recognition and Understanding, 1997*, strony 347–352, 1997.

[GR08] Giulia Garau, Steve Renals. Combining spectral representations for large-vocabulary continous speech recognition. *IEEE Transactions on audio, speech and language processing*, 16(3), Marzec 2008.

[HM03] Astrid Hagen, Andrew Morris. Recent advances in thr multi-stream hmm/ann hybrid approach to noise robust asr. *Computer speech and language*, 2003.

[HN03] Astrid Hagen, João P. Neto. Multi-stream processing using context-independent and context-dependent hybrid systems. *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, 2003.

[JH99] Li Jiang, Xuedong Huang. Unified decoding and feature representation for improved speech recognition. *Eurospeech'99*, 1999.

[MHS+05] Andreas Maier, Christian Hacker, Stefan Steidl, Elmar Nöth, Heinrich Niemann. Robust parallel speech recognition in multiple energy bands. *Pattern Recognition 27th DAGM Synopsium Vienna*, 2005.

[OBP98] Shigeki Okawa, Enrico Bocchieri, Alexandros Potamianos. Multi-band speech recognition in noisy environments. *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, 2:641–644, 1998.

[SSRS01] Rita Singh, Michael L. Seltzer, Bhiksha Raj, Richard M. Stern. Speech in noisy environments: Robust automatic segmentation, feature extraction and hypothesis combination. *Conference on Acoustics, Speech, and Signal Processing 2001*, 1, 2001.

[ZKSN07] András Zolnay, Daniil Kocharov, Ralf Schlüter, Hermann Ney. Using multiple acoustic feature sets for speech recognition. *Speech communication*, 2007.

[ZM08] Sherry Y. Zhao, Nelson Morgan. Multi-stream spectro-temporal features for robust speech recognition. *Proceedings of the 9th International Conference of the ISCA (Interspeech 2008)*, strony 898–901, 2008.

[ZSN05] András Zolnay, Ralf Schlüter, Hermann Ney. Acoustic feature combination for robust speech recognition. *Proc. IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing*, strony 457–460, 2005.

Kombinationsansätze

Wie Eingangs bereits dargestellt wurde, kann durch die Kombination von unterschiedlichen Erkennungssystemen eine Reduktion der Wortfehlerrate erreicht werden. In den Veröffentlichungen aus Fachzeitschriften und Fachtagungen werden unterschiedliche Ansätze zur Kombination verwendet, von denen einzelne Ansätze exemplarisch in der folgenden Übersicht dargestellt werden sollen, um einen kurzen Überblick über die Möglichkeiten der Kombination zu vermitteln. In den Beispielen wird die Kombination für zwei unterschiedliche Erkennungssysteme dargestellt, je nach Anwendungsfall kann diese Anzahl aber auch erweitert werden.

Kombination der akustischen Merkmale

Eine Möglichkeit zur Kombination ist die Kombination der akustischen Merkmale. Die Vorgehensweise ist schematisch als Blockschaltbild in Abbildung 1 dargestellt:

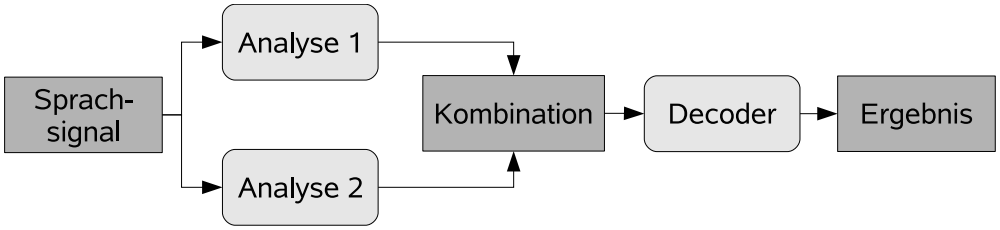

Abbildung 1: Kombination der akustischen Merkmale

Bei dieser Art der Kombination wird das Sprachsignal zwei unterschiedlichen Verarbeitungsblöcken zur Merkmalsextraktion (Analyse 1 und 2) zugeführt. Aus den zwei resultierenden Merkmalsströmen werden anschließend korrespondierende Merkmalsvektoren miteinander verbunden. Da der sich ergebende Merkmalsstrom in dieser Form noch nicht weiterverarbeitet werden kann, wird durch eine Nachverarbeitung eine Merkmals-Dekorrelation sowie eine Reduktion der Dimensionalität des Merkmalsstrom durchgeführt. Hierzu werden verschiedenen Ansätze, z.B. „Principal Component Analyse“ (PCA) oder „Linear Discriminant Analyse“ (LDA) sowie „Heteroscedastic Linear Discriminant Analyse“ (HLDA) angewendet.

In verschiedenen Fachartikeln wurde durch Anwendung dieses Kombinationsansatzes eine Reduktion der Wortfehlerrate erreicht. So stellt Burget (Bur04) in seiner Arbeit dar, dass die Verwendung von LDA und HLDA sowie daraus abgeleitete Vorgehensweisen (SHLDA und CHLDA) zu einer Reduktion der Wortfehlerrate führen, PCA hingegen stellte sich in dieser Arbeit im Vergleich als wenig brauchbar heraus. Auch Garau und Renals (GR08) konnten in ihrer Arbeit nachweisen, das es durch die Verwendung von HLDA bei der Merkmalskombination zu einer Reduktion der Wortfehlerrate kommt.

Kombination der Erkennungsergebnisse (Hypothesen Kombination)

Eine weitere Kombinationsmöglichkeit ist die Kombination der Erkennungsergebnisse verschiedener Erkennungssysteme. Der grundsätzliche Aufbau ist als Blockschaltbild in Abbildung 2 dargestellt:

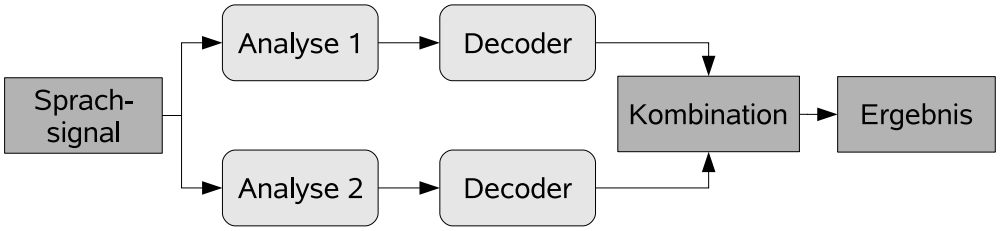

Abbildung 2: Kombination der Erkennungsergebnisse

Wie bereits zuvor dargestellt, wird das Sprachsignal zwei unterschiedlichen Verarbeitungsblöcken zur Merkmalsextraktion zugeführt. Die sich ergebenden Merkmalsströme werden bei diesem Ansatz jeweils zwei Verarbeitungsblöcken zur Erkennung (hier allgemein als „Decoder“ bezeichnet) zugeführt. Nach der getrennten Verarbeitung werden abschließend die resultierenden Ergebnisse der einzelnen Erkennungssysteme miteinander kombiniert. In dem Artikel von Singh et al. (SSRS01) wird die Kombination der Ergebnisse durch die Verwendung eines „Wortgraph“ durchgeführt. Die Wortgraph-Erzeugung aus den Ergebnissen der unterschiedlichen Erkenner erfolgt dabei mehrstufig. In der Initialstufe wird jedes Wort aus den Ergebnissen der Erkenner als Knoten abgelegt. In einer weiteren Stufe wird versucht, den Wortgraph zu optimieren. Hierbei wird z.B. hinsichtlich gleicher Ergebnisse mit ähnlichem zeitlichem Auftreten eine Zusammenfassung der Wörter vorgenommen. Nach der Konstruktion des Wortgraph wird mittels Wortmodellen eine Bewertung der Pfade des Graphen vorgenommen, der Pfad mit der höchsten Bewertung wird abschließend als Ergebnis verwendet. Die Autoren konnten bei der Anwendung des Kombinationsansatzes eine Reduktion der Wortfehlerrate verzeichnen.
Ein weiterer Ansatz zur Kombination der Erkennungsergebnisse mit dem Namen „ROVER“ stellte Fiscus in seiner Arbeit vor (Fis97). Die nach der Erkennung mit unterschiedlichen Erkennungssystemen vorliegenden Ergebnisse werden in einem „Word Transition Network“ kurz WTN abgelegt. Als Beispiel werden in der Veröffentlichung drei verschiedenen allgemeine Erkennungsergebnisse von nicht näher spezifizierten Systemen dargestellt, um die Vorgehensweise zur Konstruktion des resultierenden WTN zu erläutern (siehe Abbildung 5). Diese sind in Tabelle 15 dargestellt:

Erkenner	Ergebnis
System 1, WTN1	a,b,c,d
System 2, WTN2	b,z,d,e
System 3, WTN3	b,c,d,e,f

Tabelle 15: Beispiel-Ergebnisse ROVER

Jedes der drei Ergebnisse wird zunächst als separates lineares WTN betrachtet. Als Basis-WTN dient das Ergebnis von System 1 aus dem unter Verwendung der WTNs der Systeme 2 und 3 das resultierende WTN erzeugt wird. Aus dem Beispiel in Tabelle 15 wird deutlich, dass sich die Erkennungsergebnisse durchaus unterscheiden können. Um die WTNs aufeinander abzustimmen wird zur Anpassung der WTNs aneinander ein Verfahren zur dynamischen Programmierung verwendet. Dadurch wird eine bestmögliche Abstimmung der einzelnen Elemente der WTNs zueinander gewährleistet. Das Ergebnis der Anpassung von WTN 2 und dem Basis- WTN ist in Abbildung 3 dargestellt:

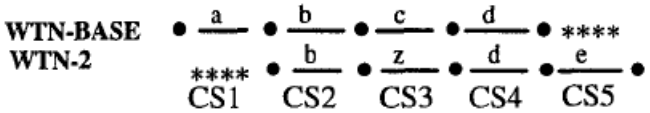

Abbildung 3: angepasste WTNs (Fis97)

Nach der Anpassung der beiden WTNs werden die Elemente (CS1 bis CS5) von WTN 2 nach festgelegten Regeln als Wortübergangsbögen („word transition arcs“) in das Basis-WTN kopiert. Hierbei stellt jedes erkannte Wort in dem Netzwerk eine Verbindung zwischen zwei Knoten dar (Hier mit den Buchstaben a, b, c etc. sowie @ für Einfügungen bzw. Auslöschungen bezeichnet):

- Regel 1: „korrektes Ergebnis“
  Eine Kopie des entsprechenden Elements aus WTN 2 wird an die Stelle des korrespondierenden Wortes des Basis- WTN hinzugefügt (CS2 und CS4 im Beispiel).

- Regel 2: „Ersetzung“
  Eine Kopie des entsprechenden Elements aus WTN 2 wird zum Basis- WTN hinzugefügt (CS3).

- Regel 3: „Auslöschung“
  Ein „no-cost“ NULL Wortübergang wird dem Basis- WTN hinzugefügt. Dieser wird mit „@“ gekennzeichnet (CS1).

- Regel 4: „Einfügung“
  Es wird ein SUB-WTN erzeugt das aus dem entsprechenden Element aus WTN 2 und einem NULL Wortübergang (mit „@“ gekennzeichnet) besteht. Dieses wird an der entsprechenden Stelle zwischen den benachbarten Knoten im Basis- WTN eingesetzt (CS5).

Das als Zwischenergebnis resultierende WTN ist in Abbildung 4 dargestellt:

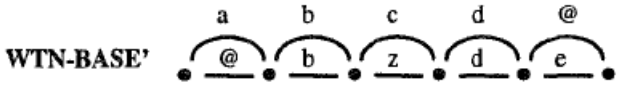

Abbildung 4: resultierendes WTN aus WTN 1 und 2 (Fis97)

In gleicher Weise wird auch mit WTN 3 verfahren. Das resultierende endgültige WTN ist in Abbildung 5 dargestellt.

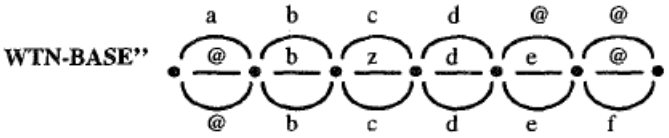

Abbildung 5: Word Transistion Network (Fis97)

Nach der Generierung des WTN wird mittels eines „Voting Search Module“ jeweils das Wort mit der höchsten Bewertung zwischen den jeweiligen Knoten des WTN ausgewählt und verwendet. Für die Bewertung sind drei verschiedenen Auswertungsverfahren vorgesehen. Hierbei werden die folgenden Methoden berücksichtigt:

1. Berechnung der Frequenz des Auftretens einzelner Wörter zwischen zwei Knoten in dem Netzwerk

2. Berechnung der Frequenz des Auftretens einzelner Wörter und Berechnung der mittleren Wortkonfidenz für jedes Wort

3. Berechnung der Frequenz des Auftretens einzelner Wörter und Berechnung der maximalen Wortkonfidenz für jedes Wort

Für die drei Auswertungsverfahren resultieren ähnliche Ergebnisse, das beste Ergebnis lieferte Verfahren 3. Auch bei der Anwendung dieses Verfahrens zur Kombination konnte eine Senkung der Wortfehlerrate beobachtet werden. In der Arbeit von Maier et al. (MHS$^+$05) wird das ROVER Verfahren zur Kombination von Erkennungsergebnissen angewendet und auch hier kann gezeigt werden, dass das Verfahren zur Senkung der Wortfehlerrate führt.

Kombination der Posterior- Wahrscheinlichkeiten

Als dritte Kombinationsmöglichkeit bietet sich die Kombination der Posterior- Wahrscheinlichkeiten an. Als Beispiel sei hier die Arbeit von Ellis (Ell00b) genannt. Das Blockschaltbild des Verfahrens ist in Abbildung 6 dargestellt.

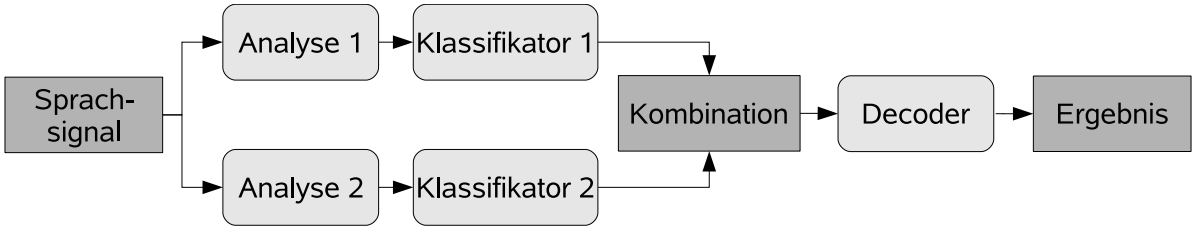

Abbildung 6: Kombination der posterior Wahrscheinlichkeiten

Hierbei wird wie bereits zuvor dargestellt, das Sprachsignal zwei unterschiedlichen Verarbeitungsblöcken zur Merkmalsextraktion zugeführt. Die sich ergebenden Merkmalsströme werden anschließend in einem Verarbeitungsblock zur Klassifizierung weiterverarbeitet. In diesem Klassifizierungsprozess werden anhand von akustischen Modellen, die z.B. als „Multilayer Perceptron Neural Networks“ vorliegen, Posterior- Wahrscheinlichkeiten für z.B. einen Satz von kontextunabhängigen Phonemen für die einzelnen Merkmalsströme berechnet. Diese werden anschließend kombiniert, indem z.B. der Mittelwert über die logarithmischen Wahrscheinlichkeiten der einzelnen Klassifikatoren gebildet wird. Abschließend erfolgt eine Konvertierung der resultierenden Posterior- Wahrscheinlichkeiten, um diese in einem Hidden- Markov- Dekoder zu verwenden. Die Anwendung des Verfahrens zur Kombination führt zu einer Verbesserung der Erkennungsrate.
Ein ähnliches Verfahren, das ebenfalls mit Posterior- Wahrscheinlichkeiten arbeitet, stellen Jiang und Huang (JH99) in ihrer Arbeit vor. Bei diesem Ansatz wird ein Verfahren zur Multiplen Merkmalsdekodierung angewendet, bei dem versucht wird, aus allen Merkmalsströmen den am besten geeigneten Pfad zu schätzen (siehe Abbildung 7, graue Schraffur).

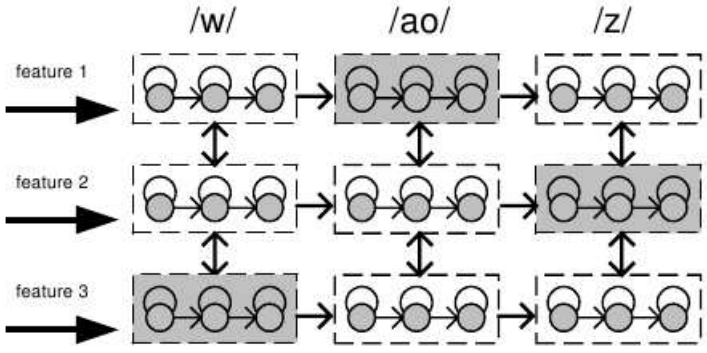

Abbildung 7: Multiple Merkmalsdekodierung (JH99)

Um eine Auswahl der Merkmalsströme zu treffen, wird eine Gewichtung der Ströme durch eine Laufzeitbewertung vorgenommen und anschließend die Wahrscheinlichkeit für den besten Pfad über alle Merkmalsströme berechnet. Als Modelle werden in dem Beispiel Phonemmodelle verwendet. Die Autoren konnten eine Senkung der Wortfehlerrate bei der Anwendung des Verfahrens verzeichnen. Im Vergleich zu den mit ROVER (Fis97) erzielten Ergebnissen konnte eine Verbesserung erzielt werden.

Kombination mittels Tandem-Verfahren

Als Erweiterung der in den vorherigen Abschnitten beschriebenen Kombinationsverfahren existieren verschiedene Ansätze, die so genannte Tandem- Verfahren verwenden. Diese bestehen aus einer Kombination der einzelnen Grundverfahren. In Abbildung 8 ist als Beispiel für ein solches Tandem- System der Ansatz von Garau und Renals (GR08) dargestellt.

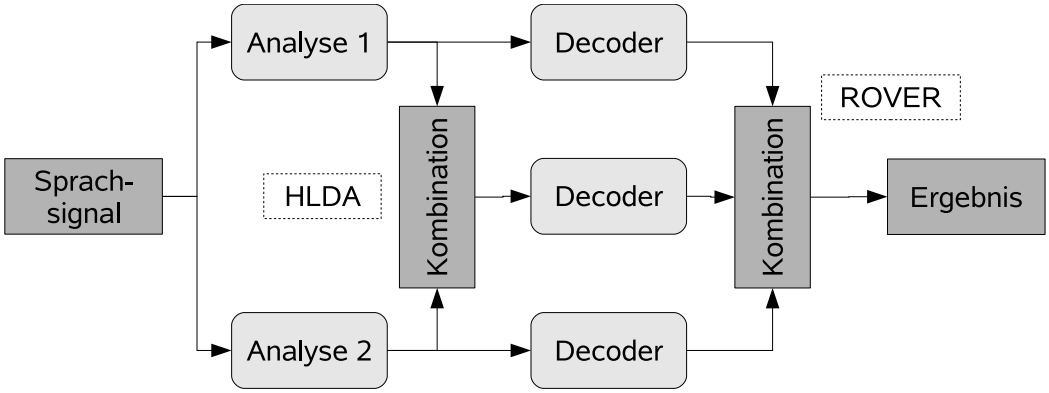

Abbildung 8: Tandem-Ansatz

Bei diesem Ansatz werden zwei verschiedene Verfahren zur Merkmalsextraktion verwendet, deren resultierende Merkmalsströme anschließend mittels „Heteroscedastic Linear Discriminant Analyse“ (HLDA) kombiniert werden. Die kombinierten Merkmale werden anschließend einem Verarbeitungsblock zur Decodierung/Erkennung zugeführt. Weiterhin werden die aus den zwei Analyseverfahren gewonnenen Merkmale direkt einem Verarbeitungsblock zur Decodierung/Erkennung zugeführt. Die Ergebnisse der drei Erkenner werden abschließend mittels ROVER (Fis97) erneut kombiniert. Die Autoren konnten nach der ersten Kombination mittels HLDA bereits eine Reduktion der Wortfehlerrate feststellen, nach der zusätzlichen Anwendung von ROVER konnte eine weitere Reduktion der Wortfehlerrate verzeichnet werden, was die Autoren zu dem Schluss brachte, dass beide Verfahren komplementär arbeiten.
Einen weiterer Tandem- Ansatz stellen Hagen und Neto (HN03) in ihrer Arbeit vor. Das als „All Combination“ bezeichnete System ist in Abbildung 9 dargestellt:

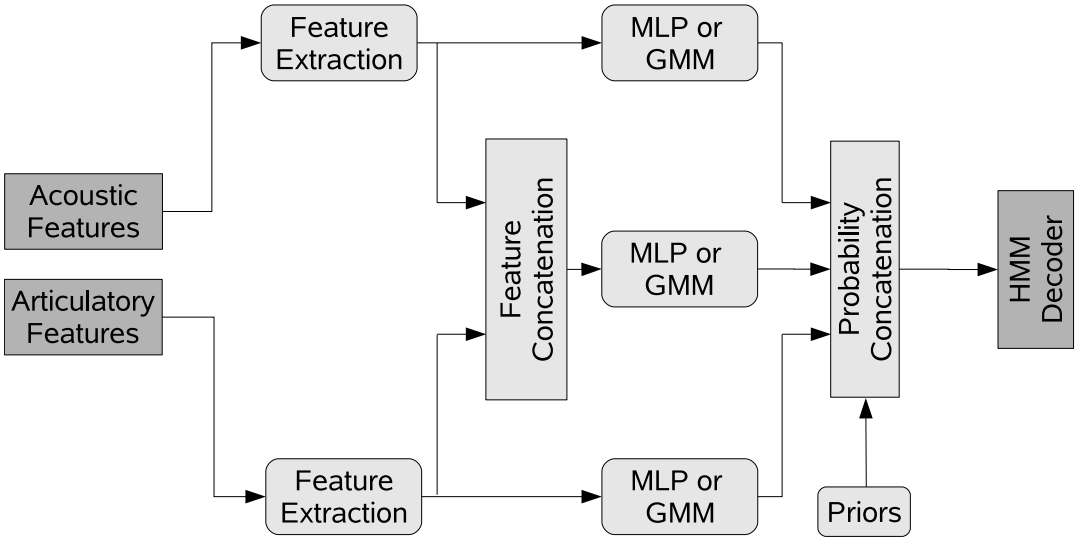

Abbildung 9: „All Combination“

Auch bei diesem Ansatz werden zwei Kombinationsverfahren miteinander verknüpft. Hierzu werden mittels zwei unterschiedlichen Verfahren zur Merkmalsextraktion zwei Ströme mit Merkmalen gewonnen. Diese werden parallel weiterverarbeitet. Beide Merkmalsströme werden zum einen in einem Verarbeitungsblock zur Merkmalskombination verarbeitet (das Verfahren zur Merkmalskombination wird nicht näher erläutert), anschließend wird aus dem kombinierten Merkmalsstrom ein Multilayer Perzepton (MLP) oder ein Gauss-Mixture Modell (je nach Ansatz) trainiert , zum anderen werden aus den einzelnen Merkmalströmen direkt MLPs oder GMMs trainiert. Die MLPs/GMMs werden anschließend einem weiteren Verarbeitungsblock zur Kombination der Posterior- Wahrscheinlichkeiten der MLPs/GMMs zugeführt. Nach der Kombination der Posterior-Wahrscheinlichkeiten werden die Ausgangswerte als skalierte Likelihoods im HMM (Hidden Markov Modell) zur Dekodierung/ Erkennung verwendet. Bei der Anwendung einer Kombinationsmöglichkeit konnten die Autoren bereits eine Senkung der Wortfehlerrate beobachten, bei der gekoppelten Anwendung von zwei Kombinationsmöglichkeiten im Tandem- Ansatz wurde die Wortfehlerrate noch weiter gesenkt.

Fazit „Kombinationsmöglichkeiten“

Im Anhang wurden die verschiedenen Verfahren zur Kombination von Erkennungssystemen vorgestellt aus denen Ideen für eigene Ansätze zur Kombination der beiden robusten Erkennungssysteme „HGH adapt“ und „HGH robust“ abgeleitet werden können. Bei der Betrachtung der verschiedenen Ansätze wird deutlich, dass die Ansätze, die neben HMMs auch MLPs verwenden als Ansatz für die Kombination der vorliegenden Systeme ungeeignet sind, da diese eine zu große Modifikation der bereits bestehenden Strukturen der

Erkennungssysteme zur Folge hätten.
Eine Kombination der Merkmale kommt wegen der verschiedenen Ansatzpunkte der Systeme ebenfalls nicht in Frage, da sich hierbei robuste und nicht robuste Merkmale wieder „vermischen“ würden, da bei dem robusten Ansatz „HGH adapt“ eine nicht robuste Cepstralanalyse ohne Störreduktion Verwendung findet.
Aus diesem Grund bietet es sich an, einen Kombinationsansatz zu verwenden, der eine Kombination der Erkennungsergebnisse (Hypothesen- Kombination) durchführt. Als Ansatz würde sich hier die Verwendung von ROVER (siehe (Fis97)) anbieten, da dieser Ansatz neben der Hauptveröffentlichung des Autors auch in weiteren Veröffentlichungen als System besprochen wurde, das eine funktionierende und sinnvolle Kombination ermöglicht. Weiterhin müssen die bestehenden Ansätze zur Verwendung von ROVER nicht modifiziert sondern nur erweitert werden.