Signalbasierte Verfahren zur robusten Spracherkennung im Cockpit von Luftfahrzeugen

Diplomarbeit
am
Institut für Logik, Komplexität und Deduktionssysteme
Interactive Systems Labs
Prof. Dr. A. Waibel
Fakultät für Informatik
Universität Karlsruhe (TH)

von

cand. inform.
Michael Dambier

Betreuer:
Prof. Dr. A. Waibel
Dipl.-Inform. Ch. Fügen

Tag der Anmeldung: 01. Juni 2003
Tag der Abgabe: 28. November 2003

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, 28. November 2003

Michael Dambier

Danksagung

Spracherkennung im Luftfahrzeug kann einen erheblichen Beitrag zur Erhöhung der Flugsicherheit leisten. Dies war mein Anliegen und meine Motivation zur Erstellung dieser Arbeit. Ich möchte mich auf diesem Wege bei Prof. Dr. Alex Waibel für die Ermöglichung der Diplomarbeit, das Interesse an dieser Arbeit und die Unterstützung besonders im Bereich der Datensammlung im Helikopter bedanken. Für die ständige Hilfe und Unterstützung bei auftretenden Problemen, Diskussionen und Verbesserungsvorschläge möchte ich meinem Betreuer Christian Fügen herzlich danken.

Danke auch an die Mitarbeiter der Interactive Systems Labs der Universität Karlsruhe (TH) Florian Metze, Dr. Ivica Rogina, Hagen Soltau, Sebastian Stüker und Matthias Wölfel für ihr Interesse an meiner Arbeit, für zahlreiche Diskussionen und Verbesserungsvorschläge.

Besonderer Dank gilt meinen Eltern, Renate und Karl-Heinz Dambier, für die Unterstützung meiner Arbeit und an den Fluglehrer Dr. Jochen Hinkelbein, der mir die Luftfahrt in all ihren Facetten näher gebracht hat und die Kosten für die Datensammlung im Sportflugzeug übernahm.

Inhaltsverzeichnis

1 Einleitung

Die ARPA[1] publizierte in den „Proceedings of the ARPA SLT Workshop" 1995 eine „Vorstellung" (Abbildung 1.1), nach der im Jahr 2002 Spracherkennung mit Sprechern mit verschiedenen Sprechstilen in verschiedenen geräuschbehafteten Umgebungen, im Idealfall überall, möglich sein sollte. Dies ist heutzutage sehr wohl mit auf diesen Geräuschumgebungen trainierten Spracherkennern möglich. Leider gibt es manche Anwendungen, in denen die Kosten ein Training mit originalen geräuschbehafteten Daten unmöglich machen, so dass auf Akustiken reiner Sprache zurückgegriffen werden muss. In dieser Arbeit wurden nun im Jahr 2003 einige signalbasierte Verfahren zur Geräuschreduktion auf ihren Einsatz in der Spracherkennung in verschiedenen Luftfahrzeugen der Allgemeinen Luftfahrt mit einem Spracherkennungssystem mit „Laborakustik" überprüft.

1.1 Motivation

Piloten müssen bei ihrer Arbeit vielfältige Aufgaben im Cockpit bewältigen. Diese Mehrarbeit, vor allem in außergewöhnlichen Situationen, stellt eine erhebliche Belastung für Cockpitbesatzungen dar und führte in seltenen Fällen sogar zu Flugunfällen. Der Einsatz von Spracherkennungstechnologie im Cockpit zur Eingabe von für den Flug benötigten Parametern kann dieses Risiko mindern und den Piloten entlasten. Dabei darf das System keinesfalls auf primäre, die Konfiguration (Flugzustand) des Luftfahrzeugs bestimmende Elemente Einfluss nehmen. Das Spracherkennungssystem sollte eine Erleichterung bei der Bedienung des GPS[2], des FMS[3], der Einstellung von Kommunikations- und Navigationsfrequenzen und der Einstellung von Sekundär-Radar-Codes bieten. Gerade im unkontrollierten Sichtflugverkehr, der eine ständige Beobachtung des Luftraums erfordert, ist ein Einstellen der genannten Geräte mit einer kurzzeitigen Sichtunterbrechung nach draussen verbunden. Dies ist bei hohen Fluggeschwindigkeiten ein erhebliches Sicherheitsrisiko. Mit der Bereitstellung eines verlässlichen Spracherkennungssystems kann das Einstellen verschiedener Parameter ohne Sichtunterbrechung erfolgen und die Flugsicherheit

[1] Advanced Research Projects Agency
[2] Global Positioning System
[3] Flight Management System

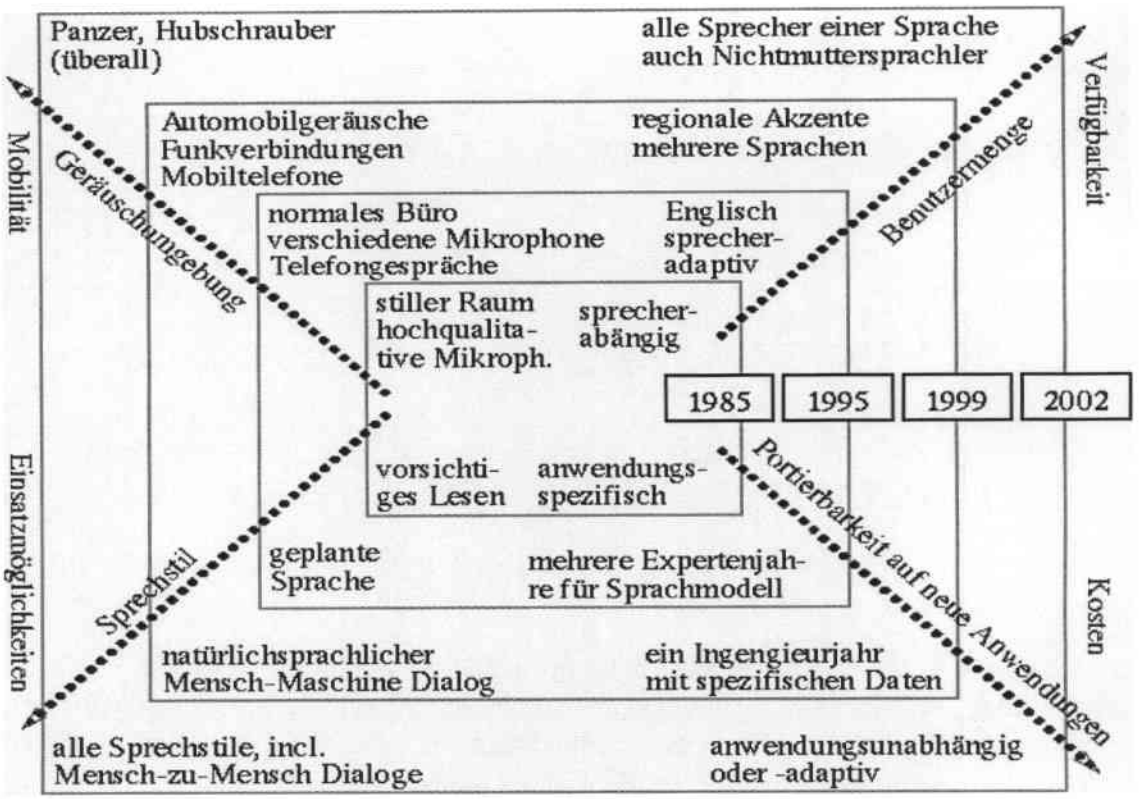


Abbildung 1.1: Vorstellung der ARPA (SLT Workshop 1995)

damit erhöht werden. Durch eine korrekte Funktion können auch Einstellungsfehler der Flugzeugbesatzung, welche unbemerkt zu einem Flugunfall (CFIT[4]) führen können, vermieden werden.

Weiterhin werden komplexe Arbeitsabläufe abgekürzt. So ist beispielsweise bei einem notwendigen Wechsel auf eine andere Funkfrequenz zuerst ein Nachschlagen der benötigten Frequenz in den Flugunterlagen und danach ein Einstellen dieser an einem Funkgerät notwendig. Diese Abläufe können durch ein Spracherkennungssystem nach dem Kommando „Change frequency to Munich tower" selbständig durchgeführt werden. Auch ein Umplanen der Flugroute im GPS ist eine komplexe Tätigkeit, welche viele Tastatureingaben und Menüwechsel erfordert. Durch ein Spracherkennungssystem kann auch diese Tätigkeit mit wenigen Kommandos durchgeführt werden. Der Pilot hat damit wiederum mehr Zeit, sich seiner eigentlichen Aufgabe, dem Führen des Luftfahrzeuges, zu widmen.

1.2 Zielsetzung der Arbeit

Zur Entwicklung einer entsprechenden Akustik für einen Spracherkenner im Luftfahrzeugcockpit fallen, wie bereits erwähnt, hohe Kosten an, da mehrere Stunden Sprachdaten aus dem Cockpit eines fliegenden Luftfahrzeuges benötigt werden und die Preise pro Flugstunde sehr hoch sind. Um diese Kosten zu vermeiden, kann auf bereits existierende „Laborakustiken", Akustiken störungsfreier Sprache oder mit anderen Geräuschen gestörter Sprache, zum Beispiel Lärm eines fahrenden Autos,

[4]Controlled Flight Into Terrain

zurückgegriffen werden. Die Erkennungsraten derartiger Laborsysteme verschlechtern sich aber zunehmend mit steigender Geräuschkulisse.

Ziel dieser Arbeit ist es, verschiedene Vorverarbeitungsverfahren für eine Verwendung in einem Spracherkennungssystem im Cockpit eines Luftfahrzeuges zu überprüfen und ein Verfahren zu finden, welches die Verwendung eines Spracherkenners mit „Laborakustik“ in einem Luftfahrzeugcockpit mit zufriedenstellendem Ergebnis ermöglicht. Aufgrund der großen Anzahl möglicher Verfahren wurden hier nur wenige betrachtet und versucht, diese durch eine verbesserte Geräuschschätzung und Kanalkompensation mit Hilfe einer adaptiven bereichskombinierten Sprach-Pause-Detektion für einen Einsatz verwendbar zu machen. Zur Verbesserung der Spracherkennungsraten sollten keine künstlich verrauschten Daten zum Training eines neuen Spracherkenners benutzt werden, da dieses Vorgehen bereits erfolgreich angewendet wurde. Ein Beispiel ist in [West01] für die Spracherkennung im Auto zu finden.

1.3 Gliederung der Arbeit

Im folgenden Kapitel wird der aktuelle Stand der Technik dargestellt. Nach einer Erläuterung der Lärmsituation und der Vorstellung bereits in Luftfahrzeugen eingesetzter Spracherkenner folgt eine Abgrenzung zu anderen Arbeiten.

Im Kapitel 3 werden verschiedene Verfahren zur Geräuschreduktion erläutert. Auch die für alle Verfahren notwendige Sprach-Pause-Detektion wird ausführlich betrachtet. Es wird weiterhin eine erste Bewertung über die Verwendbarkeit der Verfahren gegeben.

Der verwendete Spracherkenner, die Trainingsdaten, die Sprachmodelle und die Sprachdaten werden in Kapitel 4 erläutert. Es werden die in den Experimenten verwendeten Gütemaße vorgestellt. Eine Darstellung der durchgeführten Experimente und eine Erläuterung der Ergebnisse befindet sich ebenfalls im Kapitel 4.

Danach erfolgt im Kapitel 5 eine Zusammenfassung der Ergebnisse dieser Arbeit und ein Ausblick in die Zukunft der Spracherkennung in Luftfahrzeugen der Allgemeinen Luftfahrt.

Im Anhang A werden die Ergebnisse der im Rahmen dieser Arbeit durchgeführten Umfrage und die Datensammlung sowie die zur Datensammlung verwendeten Luftfahrzeuge und die verwendeten Anfragen an ein Flugnavigationssystem beschrieben.

2 Stand der Technik

Forschung in dem Bereich der Spracherkennungssysteme in Luftfahrzeugen wird seit 1982 vor allem von militärischen Einrichtungen betrieben. Neben dem Einsatz der Systeme in Flugzeugen stellt der Einsatz in Hubschraubern einen weiteren, dringlicheren Forschungsbereich dar. Im Gegensatz zum Flugzeug benötigt der Pilot beide Hände zur Steuerung des Hubschraubers. Somit hat er keine Hand für andere Tätigkeiten, wie zum Beispiel das Einstellen einer Funkfrequenz, frei. Es ist daher sehr sinnvoll, dem Piloten alternative Eingabemöglichkeiten zur Verfügung zu stellen. Da sprecherabhängige Systeme zur Zeit eine höhere Erkennungsrate in Geräuschumgebungen besitzen und in der militärischen Luftfahrt nur wenige Piloten ein Luftfahrzeug führen, werden diese überwiegend eingesetzt. Ein entsprechendes Training zur Anpassung der Akustik und der Modelle des Spracherkenners an den jeweiligen Sprecher kann in diesem Fall in Kauf genommen werden. Diese Systeme haben nach [Rood98b] Satzerkennungsraten von circa 97 Prozent bei einem relativ kleinen Vokabular. Dies ist durch den Kommandocharakter der Spracheingaben bedingt, wobei die „Eingabesätze" meist weniger als drei Worte umfassen.

In diesem Kapitel wird zunächst die Lärmsituation in Luftfahrzeugen beschrieben. Danach folgt eine eingehende Betrachtung von Spracherkennungssystemen in Luftfahrzeugen und eine Abgrenzung dieser Arbeit von anderen veröffentlichten Arbeiten.

2.1 Lärmsituation in Luftfahrzeugen

Luftfahrzeuge haben je nach Art – ein- oder mehrmotorig – und Lage des Antriebes – in der Zelle oder an den Tragflächen – eine sehr unterschiedliche Geräuschkulisse im Cockpit. Hauptursachen für Lärm im Cockpit sind das Triebwerk, der Propeller und die Luftströmungen an Rumpf und Tragfläche. Während bei durch Kolbenmotoren angetriebenen Flugzeugen ein eher niederfrequenter Geräuschpegel im Cockpit auftritt, sind bei Jets hochfrequente Geräusche aufgrund der aerodynamischen Strömung und der hohen Drehzahl der Turbinen vorhanden. Helikopter besitzen zusätzlich zu den Strömungsgeräuschen und den Geräuschen der Turbine oder des Kolbenmotors eine weitere Lärmquelle durch den Hauptrotor.

Während im Cockpit eines Jets Geräusche mit einem Lärmpegel von etwa 60 bis 88 dB und bei einem einmotorigen Flugzeug Geräuschpegel zwischen 70 und 90 dB auftreten, sind bei einem Helikopter Geräuschpegel zwischen 80 und 106 dB zu erwarten. Des weiteren tragen auch Warnsignale, wie zum Beispiel die Überziehwarnung beim Flugzeug oder Turbinenwarnsignale im Helikopter, zur Erhöhung der Geräuschkulisse bei. Da es in der Luftfahrzeugzelle an unterschiedlichen Stellen zu Auslöschungen oder Verstärkungen der Innengeräusche kommt, kann für verschiedene Sprecher auf unterschiedlichen Sitzpositionen eine völlig unterschiedliche Geräuschsituation bestehen. Einzige Ausnahme bildet das Segelflugzeug, das meist nur einen oder zwei Sitze besitzt und dessen Innengeräusche im Cockpit sich auf Strömungsgeräusche beschränken.

Piloten der Allgemeinen Luftfahrt tragen aufgrund der Lärmsituation im Cockpit Headsets, die einen Teil der Cockpitgeräusche dämpfen beziehungsweise unterdrücken. Der Mensch passt seine Aussprache durch Veränderungen der Stimme und der Sprechweise an den Umgebungslärm an. Dies wird als Lombard-Effekt bezeichnet und hat Auswirkungen auf die Erkennungsleistung des Spracherkenners. Die erlebte Lärmsituation entspricht nun bei Piloten, die ein Headset tragen, nicht mehr der vorhandenen Lärmsituation, so dass die Aussprache nur ungenügend angepasst wird. Dies hat ebenfalls Auswirkungen auf die Spracherkennung. Weiterhin können unterschiedliche Beschleunigungen die Aussprache des Piloten verändern. Beide Aspekte werden jedoch in der vorliegenden Arbeit nicht weiter betrachtet.

Die für diese Arbeit zur Verfügung stehenden Sprachaufnahmen wurden im Hinblick auf die störenden Geräusche analysiert. Als Analyse-Kriterien wurden das Signal-zu-Rausch-Verhältnis (SNR), die Frequenzlokalisation der maximalen mittleren Geräuschenergie sowie die Varianz des Störgeräuschs verwendet.

Das Signal-zu-Rausch-Verhältnis (SNR) gibt an, wie stark ein Signal verrauscht ist. Da die gestörten Sprachdaten reale Geräuschaufnahmen sind und nicht künstlich gemischt wurden, wurde zur Berechnung des SNR zunächst mit einem Sprach-Pause-Detektor die Bereiche reiner Geräusche und die Bereiche der mit Geräusch gestörten Sprache berechnet. Auf diesen beiden Bereichen wurde die mittlere Energie bestimmt – die mittlere Energie des gestörten Sprachsignals s^{pow} sowie die mittlere Energie der Sprachpausen n^{pow} – und das SNR berechnet. Dies ist unter der Annahme eines für die Dauer der Aufnahme stationären Rauschens mit zeitlich gleichbleibenden statistischen Eigenschaften möglich [West01].

$$SNR = 10 \cdot log_{10} \frac{\left(s^{pow} - n^{pow}\right)^2}{\left(n^{pow}\right)^2} \; dB \tag{2.1}$$

mit

$$s^{pow} = \frac{1}{\sum_{k=0}^{N-1} d_{Sprache}[k]} \cdot \sum_{k=0}^{N-1} d_{Sprache}[k] \cdot s^{pow}[k] \tag{2.2}$$

und

$$n^{pow} = \frac{1}{\sum_{k=0}^{N-1} d_{Pause}[k]} \cdot \sum_{k=0}^{N-1} d_{Pause}[k] \cdot n^{pow}[k] \tag{2.3}$$

wobei $d[k]$ als Ausgabe des Sprach-Pause-Detektors je nach Berechnung von s^{pow} oder n^{pow} die Werte 0 und 1 für Sprach- beziehungsweise Pauserahmen annimmt und N die Anzahl der Analyserahmen ist.

Im Folgenden werden die Ergebnisse der Analyse der einzelnen Störgeräusche – aufgenommen mit verschiedenen Mikrofonen – widergegeben. Die Spektren der Störgeräusche sowie deren Varianz sind im Anhang A.4 abgebildet.

- Cockpitgeräusche im Helikopter, aufgenommen mit einem Standard-Nahbesprechungsmikrofon der Firma Sennheiser HD-440-6

  Das Spektrum wurde über alle Aufnahmen der Gesamtlänge von circa 3,5 Minuten ermittelt. Die Aufnahmen mit einem Nahbesprechungsmikrofon zeigen einen sehr hohen Störgeräuschanteil im Bereich von 0 bis 500 Hz. Die Varianz des Störgeräuschs ist in diesem Bereich ebenfalls sehr hoch. Das SNR ist aufgrund der nicht vorhandenen Störgeräuschdämpfung des Mikrofons mit 3,4 dB sehr niedrig. Die Varianz des SNR liegt bei 2,5. Labormikrofone können aufgrund der fehlenden Störgeräuschkompensation nicht für die Sprachaufnahme in Luftfahrzeugen eingesetzt werden. Es muss auf Luftfahrt-Headsets und die dort eingebauten Mikrofone zurückgegriffen werden, da diese bereits eine Vorverstärkung und eine auf das Luftfahrzeug abgestimmte Geräuschunterdrückung besitzen.

- Cockpitgeräusche im Helikopter, aufgenommen mit einem Nahbesprechungsmikrofon des Headsets Bose Typ AH-TC

  Das Headset ist für den Einsatz in Hubschraubern konzipiert und wirkt entsprechend geräuschunterdrückend. Das Störgeräuschspektrum zeigt zwei Spitzen im Bereich von 0 bis 500 Hz und von 500 bis 1000 Hz, wobei die erste sehr hoch ausfällt, in ihrem Wert aber unter der Spitze des Labormikrofons HD-440-6 liegt. Dies zeigt die geräuschdämpfende Eigenschaft des Headset-Mikrofons. Die größte Varianz des Störgeräuschs ist zwischen 0 und 400 Hz lokalisiert. Das SNR liegt bei 10,4 dB, die Varianz bei 10,6.

- Cockpitgeräusche im Helikopter, aufgenommen mit einem Nahbesprechungsmikrofon des Headsets Sennheiser HME 100

  Das HME 100 ist ein Headset, das für den Einsatz in Sportflugzeugen entwickelt wurde. Es dämpft daher die tiefen Störgeräusche des Kolbenmotors. Zur besseren Vergleichbarkeit der Testergebnisse wurden mit diesem Headset Aufnahmen im Helikopter gemacht. Dabei lag das SNR der Aufnahmen bei 10,9 dB und ist somit nicht deutlich schlechter als die Aufnahmen, die mit dem Bose-Headset gemacht wurden. Die Varianz des SNR ist mit 2,9 aber deutlich geringer. Die maximale mittlere Geräuschenergie ist im Bereich von 200 bis 900 Hz lokalisiert, was die Dämpfungseigenschaften des Headsets zeigt. Allerdings sind im Bereich von 1000 bis 5000 Hz weitere Spitzen vorhanden, die das Spektrum der Aufnahmen mit dem Bose-Headset nicht aufweist. Weiterhin ist auch der maximale Wert des Störgeräuschs deutlich höher. Die maximale Varianz liegt um 500 Hz.

- Cockpitgeräusche im Sportflugzeug, aufgenommen mit einem Nahbesprechungsmikrofon des Headsets Sennheiser HME 100

  Das durchschnittliche Störgeräusch aller Aufnahmen mit einer Dauer von circa 49 Minuten zeigt die maximale Energie zwischen 0 und 300 Hz. Die größte Varianz des Störgeräuschs liegt ebenfalls in diesem Bereich. Das SNR ist 5,75 dB,

die Varianz 51,0. Die sehr große Varianz des SNR hat viele verschiedene Störgeräusche in unterschiedlichen Flugphasen – Rollen, Start, Steigflug, Reiseflug, Sinkflug und Landung – als Ursache. Beispielsweise klapperte beim Rollen am Boden die Sonnenschutzblende aufgrund von Vibrationen sehr stark. Im Flug trat dies dann nicht mehr auf.

Die menschliche Sprache ist zwischen den Frequenzen 100 Hz und 10 kHz lokalisiert, wobei die meiste Energie im Frequenzband von 200 Hz bis 4 kHz zu finden ist. Da dieser Bereich bei den Aufnahmen aus dem Helikopter stark überdeckt wird, können bei diesen Aufnahmen schlechtere Ergebnisse als bei den Aufnahmen aus dem Sportflugzeug, bei denen der Bereich weniger überdeckt wird, erwartet werden. Die Ursache der Überdeckung bei den Helikopter-Aufnahmen ist in den Geräuschen der Turbine zu sehen, da diese Störgeräusche in mittleren bis hohen Frequenzbereichen emitiert. Allerdings muss bedacht werden, dass die maximale Störgeräuschenergie beim Sportflugzeug höher ist als die des Helikopters. Dies kann zu einer stärkeren Verschlechterung der Ergebnisse führen als die überdeckten Frequenzbereiche der Sprache.

2.2 Spracherkenner in Luftfahrzeugen

In diesem Abschnitt werden unterschiedliche Spracherkennungssysteme in Luftfahrzeugen beschrieben. Die Darstellung ist keineswegs vollständig. Da Forschung vor allem im Bereich der Spracherkennung in Luftfahrzeugen von militärischen Einrichtungen betrieben wird, sind nicht alle Ergebnisse öffentlich. Diese Darstellung stellt nur einen kleinen Teil der öffentlich zugänglichen Forschungsdaten dar.

Das niederländische National Aerospace Laboratory NLR hat in Zusammenarbeit mit Philips im Projekt FalconEar eine Spracherkennung für den Piloten eines F-16 Kampfflugzeuges erstellt [NLR02]. Dabei kam der kommerzielle von Philips vertriebene Spracherkenner VoCon auf einem Standard-PC mit dem Betriebssystem Windows 2000 zum Einsatz. Der Spracherkenner gewährleistet Sprecherunabhängigkeit. Die Spracheingabe erfolgt kommandobasiert. Aufgrund dessen ist nur ein kleines Vokabular notwendig, das abhängig von der jeweiligen Flugphase gewechselt wird. Die Angabe der Vokabulargröße sowie der Erkennungsrate fehlt. Der Spracherkenner wurde im Simulator evaluiert. Die Aufnahme der Sprache des Piloten erfolgt über ein an der Sauerstoffmaske eingebautes Mikrofon. Hierdurch besteht eine ganz andere Geräuschkulisse als bei der Aufnahme mit einem Headset, da durch die Sauerstoffmaske (Abbildung 2.1) die Cockpitgeräusche zusätzlich gedämpft werden. Eine Wort- beziehungsweise Satzerkennungsrate wurde nicht angegeben. Ein weiteres Projekt des NLR mit einem kommandobasierten Spracherkenner für den militärischen Einsatz ist in [Offe97] beschrieben.

In [Will97] wird ein kommerzieller militärischer Spracherkenner beschrieben, der mit im Flugzeug aufgenommenen Sprachdaten trainiert wurde. Er ist sprecherabhängig, arbeitet kommandobasiert und besitzt daher ein kleines Vokabular mit 53 Wörtern. Getestet wurde dieser Spracherkenner in einer zweimotorigen Turboprop. Erreicht wurden Worterkennungsraten von 98 Prozent.

[SwKo97] beschreibt einen sprecherabhängigen, kommandobasierten Spracherkenner für Helikopter. Dieser ist als DVI[1]-Karte in das CMA-2082 Avionics Management

[1] Direct Voice Input

Abbildung 2.1: Helm und Sauerstoffmaske eines Jetpiloten

System der Firma Canadian Marconi Company eingebaut. Der Spracherkenner verwendet eine kontinuierliche Spracherkennung und reagiert auf festgelegte Schlüsselwörter. Diese Schlüsselwörter muss der Pilot lernen und beherrschen, um das System auch in Stresssituationen bedienen zu können. Damit der Pilot dies leisten kann, beschränkt sich das Vokabular auf 80 Wörter. Mit diesem System wurden Worterkennungsraten von 94,9 Prozent erzielt. Es besteht die Möglichkeit der Erweiterung des Vokabulars auf 800 Wörter, die nicht genutzt wurde, da der Pilot dann eine komplexe Anfragen-Syntax lernen und beherrschen muss.

Ein weiterer sprecherabhängiger Spracherkenner mit kontinuierlicher Erkennung wird in [WiBL96] beschrieben. Er ist kommandobasiert und besitzt ein Vokabular von 54 Wörtern. Im Gegensatz zu den anderen vorgestellten Projekten besitzt dieses System ein weiteres Mikrofon zur Aufnahme der Umgebungsgeräusche im Cockpit. Damit wurden mittlere Worterkennungsraten von ungefähr 98 Prozent erzielt.

Eine weitere Arbeit im Bereich der Spracherkennung in Luftfahrzeugen beschreibt Gerlach in [Gerl96]. An der Bundeswehruniversität in München wurde ein sprachgesteuertes Cockpitassistenzsystem entwickelt. Der Spracherkenner ist sprecherunabhängig und es wird eine kontinuierliche Spracherkennung verwendet. Das Vokabular wurde nach den Regeln des Funkverkehrs und denen des Crew Coordination Concepts erstellt. Allerdings werden wiederum Kommandos beziehungsweise Referenzsätze zur Spracherkennung verwendet. Je nach der Phase (Start, Reiseflug, Landung etc.) des Fluges erfolgt eine kontextabhängige Einschränkung des Sprachmodells. Verwendet wird der Verbundworterkenner PE/DS200 der Firma Speech Systems Incorporated auf einer SUN Sparc Station SS-20. Mehrere Experimente wurden mit unterschiedlichen Vokabularen durchgeführt. Die durchschnittliche Erkennungsrate ist mit 85,5%, die durchschnittliche Vokabulargröße mit 94 Wörtern angegeben. Die Spracherkennung wird zur Ansteuerung des Flight Management Systems sowie des zusätzlich entwickelten Cockpitassistenzsystems verwendet. Das System ist nur für Berufspiloten und für den Einsatz beim Instrumentenflug geeignet. Es wurde in einem Simulator getestet. Eine Robustheit gegenüber realen Cockpitgeräuschen wurde nicht nachgewiesen.

Im Eurofighter Typhoon [Euro01] sowie im Joint Strike Fighter von Boeing [Tull00] kommen ebenfalls Spracherkenner für den „Direct Voice Input (DVI)" zum Einsatz. Dabei verwendet Boeing den kommerziellen kontinuierlichen Spracherkenner Voxware der Firma ITT Industries. Im Eurofighter kommt ein sprecherabhängiges Hardware-Spracherkennungsmodul der Firma Smith Industries zum Einsatz. Dieses Modul wurde für kontinuierliche Spracherkennung in einer geräuschbehafteten Cockpitumgebung und für die Erkennung von Sprache unter hohen Beschleunigungsbelastungen des Piloten ausgelegt. Die Spracherkennung erfolgt mit Hilfe von Markov-Modellen und neuronalen Netzen. Mit einem Vokabular von 200 Wörtern und einer durchschnittlichen Antwortzeit des Systems von 120 ms bei einer Worteingaberate von 120 bis 140 Wörtern pro Minute wird die „Erkennungsfähigkeit" mit mehr als 95 Prozent angegeben. Die Angabe einer Worterkennungsrate fehlt. Vor der ersten Benutzung des Systems muss jeder Pilot in weniger als einer Stunde das System durch Sprachtraining auf sich adaptieren. Dies geschieht an sogenannten PC based Ground Support Stations. Die so trainierte Akustik wird dann vor dem Einsatz auf das System im Flugzeug überspielt. Das System ist an die Communication and Audio Management Unit (CAMU) des Jets angeschlossen. Somit wird die Sprache durch das Mikrofon in der Sauerstoffmaske des Piloten aufgenommen. Wie bereits oben beschrieben, wird dadurch eine von den realen Cockpitgeräuschen unterschiedliche Geräuschkulisse aufgenommen. Das System wird für die Steuerung von 26 unkritischen Cockpitfunktionen, wie zum Beispiel die Belegung der Bildschirme und die Auswahl von Navigationspunkten, verwendet.

2.3 Abgrenzung zu anderen Arbeiten

Im vorhergehenden Abschnitt wurden verschiedene Arbeiten aus dem Bereich der Spracherkennung in Luftfahrzeugen vorgestellt. Auffällig ist, dass viele Systeme sprecherabhängig sind und nur Kommandowörter verarbeiten können. Dies ist durchaus sinnvoll, da sprecherabhängige Systeme mit einem sehr beschränkten Vokabular gute Wort- und Satzerkennungsraten aufweisen. Es gibt auch einige sprecherunabhängige Systeme, die wiederum ein Mikrofon in der Sauerstoffmaske des Piloten nutzen. Die Sauerstoffmaske dämpft aufgrund ihrer Bauart die Cockpitgeräusche. Piloten von Helikoptern verwenden in ihre Helme eingebaute Mikrofone. Diese entsprechen den Mikrofonen der Headsets der Allgemeinen Luftfahrt. Es wurde während der Recherche kein System gefunden, welches für mehrere Luftfahrzeugarten eingesetzt werden kann und die Sprache mittels eines – in der Allgemeinen Luftfahrt gebräuchlichen – Headsets aufnimmt und eine sprecherunabhänige Erkennung spontan gesprochener Sätze ermöglicht.

Spracherkennung im Auto ist heutzutage auf dem Vormarsch. Durch diese Technologie kann der Fahrer verschiedene Informationen erfragen oder Einstellungen vornehmen, ohne dabei seine eigentliche Aufgabe, nämlich das Lenken des Kraftfahrzeugs, zu vernachlässigen. Da ein Pilot gegenüber einem Autofahrer ein erhebliches Maß an Mehrarbeit zu leisten hat, stellt sich die Frage, ob die im Auto angewandte Spracherkennung nicht auch in einem Flugzeug eingesetzt werden kann. Diese kann nur nach einer Modifikation der Signalvorverarbeitung verwendet werden, da sich die Geräuschsituation und die Aufnahmemöglichkeiten deutlich unterscheiden. Moderne Fahrgasträume von Kraftfahrzeugen sind mit viel Komfort ausgestattet. Hierzu gehört eine adäquate Geräuschreduzierung. Die verbleibenden Innengeräusche be-

tragen etwa 40 bis 73 dB(A). Eine Sprachaufnahme mit einem Headset mit Nahbesprechungsmikrofon verbietet schon die Bequemlichkeit des Fahrers und die Straßenverkehrsordnung. Deshalb werden Fernbesprechungsmikrofone mit einem - im Vergleich zum geräuschreduzierenden Nahbesprechungsmikrofon des Luftfahrtheadsets - großen Übertragungsbereich (30 Hz - 20 kHz) verwendet. Dies macht eine Modifikation der Vorverarbeitung der Sprachsignale notwendig. In dieser Arbeit wurde dies mit der modellkombinationsbasierten akustischen Transformation (MAM) versucht. Die MAM wurde ausgewählt, da Westphal sie in [West01] primär für ein Spracherkennungssystem im Auto entwickelt, aber auch für den Einsatz in anderen Geräuschumgebungen vorgeschlagen hat.

3 Signalbasierte Verfahren zur robusten Spracherkennung

Spracherkennung im Luftfahrzeug hat als Hauptproblem die sehr stark störende Geräuschkulisse. Zahlreiche andere Anwendungen besitzen dieses Problem auch, so dass der Versuch, diese zu beseitigen beziehungsweise zu minimieren, in zahlreichen, häufig signalbasierten Verfahren zur Geräuschreduktion endet. Diese sind ein Muß für eine gute Spracherkennung im Luftfahrzeug.

In diesem Kapitel sollen nach einer Betrachtung der Standard-Signal-Vorverarbeitung eines Spracherkenners einzelne Verfahren zur Geräuschreduktion näher betrachtet werden. Die beschriebenen Verfahren wurden im Rahmen dieser Arbeit für eine Verwendung in einem Luftfahrzeug-Spracherkennungssystem mit Sprachdaten aus einem Helikopter und einem Sportflugzeug getestet.

Verfahren der Geräuschreduktion wurden bevorzugt, da nach [West01] dadurch die Varianz der Modelle klein gehalten wird, um auch bei ungestörten Aufnahmen gute Erkennungsergebnisse zu erzielen. Dadurch ist es auch möglich, bereits vorhandene, trainierte „Laborakustiken" zu verwenden.

Betrachtet werden die spektrale Subtraktion, die modellkombinationsbasierte akustische Transformation und die cepstrale Mittelwertsubtraktion. Weitere Verfahren wie beispielsweise RASTA oder GDCR[1] werden nicht betrachtet, da in [dVBo96] gezeigt wurde, dass die cepstrale Mittelwertsubtraktion (CMS) diese kanalnormalisierenden Verfahren übertrifft. Weiterhin wurde in [dVBo96] festgestellt, dass die Performance des phase-corrected RASTA mit der der CMS übereinstimmt. Deshalb wurde in dieser Arbeit die Kanalkompensation nur mit Hilfe der CMS betrachtet.

Da die betrachteten Verfahren bereits in anderen Geräuschumgebungen gute Ergebnisse gezeigt haben, wurde im Rahmen dieser Arbeit versucht, durch eine verbesserte Geräuschschätzung beziehungsweise Mittelwertsubtraktion mit Hilfe einer adaptiven und damit luftfahrzeug- beziehungsweise störgeräuschunabhängigen Sprach-Pause-Detektion die Erkennungsergebnisse auf den Luftfahrzeugdaten zu verbessern. Hierbei wurde auch die Möglichkeit einer separaten Störgeräuschaufnahme mit einem

[1] Gaussian Dynamic Cepstrum Representation

zweiten baugleichen Headset-Mikrofon für die spektrale Subtraktion sowie die modellkombinationsbasierte akustische Transformation betrachtet.

3.1 Standard-Vorverarbeitung

Die Vorverarbeitung ist die erste Stufe des Spracherkennungsprozesses. Als Basis für den Vorverarbeitungsprozeß steht die Sprachaufnahme als digitales Zeitsignal $s(t)$ mit einer Abtastrate von 16 bzw. 8 kHz und einer Quantisierung von 16 bit zur Verfügung. Aus diesem Signal wird in 12 Schritten ein akustischer Merkmalsvektor extrahiert, der alle wichtigen Sprachinformationen enthält und für den eigentlichen Erkennungsprozeß verwendet wird.

Im Folgenden werden die einzelnen Stufen des Vorverarbeitungsprozesses erläutert, wie sie auch in [Wölf03] beschrieben werden. Eine Übersicht über die Stufen gibt Abbildung 3.1.

1. Segmentierung
   Sprache wird als quasi-stationäres Signal betrachtet. Die Kanaleigenschaften des Vokaltraktes können in einem Zeitraum von 5 bis 30 Millisekunden als konstant betrachtet werden. Aus diesem Grund wird das Sprachsignal zur Weiterverarbeitung in kurze Stücke (Frames) segmentiert. Hierzu wird ein Analyse-Fenster mit einer festen Breite von 16 oder 20 Millisekunden alle 10 Millisekunden (shift) mit dem Sprachsignal multipliziert. Als Analyse-Fenster wird das Hamming-Fenster benutzt.

2. Fourier-Transformation
   Auf den durch die Segmentierung erhaltenen Frames wird die schnelle Fourier-Transformation (FFT) berechnet und so das Signal in den Spektralbereich transformiert.

3. Sprach-Pause-Detektion
   Für die in einem späteren Schritt durchgeführte Kanalkompensation werden die Sprachsegmente des Signals benötigt, die durch einen Sprach-Pause-Detektor bestimmt werden. Diese Detektion wird mit einem einfachen energiebasierten Sprach-Pause-Detektor mit nicht adaptivem, also konstanten Schwellwert durchgeführt.

4. Leistungsspektrum
   Um aus dem Ergebnis der schnellen Fourier-Transformation das Leistungsspektrum zu erhalten, wird der Betrag des Ergebnisses der Fourier-Transformation quadriert. Das Leistungsspektrum wird deshalb auch als Betragsquadratspektrum bezeichnet.

5. Vokaltraktlängennormierung (VTLN)
   Verschiedene Sprecher besitzen aufgrund ihrer Größe und ihrer Physiologie verschieden lange Vokaltrakte. Um diese Unterschiede zu verringern wird die Vokaltraktlängennormierung durchgeführt. Diese verschiebt durch eine stückweise lineare Abbildung die Formanten des Sprechers in Richtung der Formanten eines „Standardsprechers". Es ist somit ein Adaptionsverfahren im Spektralbereich. Der zur Adaption notwendige Verzerrungsfaktor wird initial mit $1,0$ angenommen und während des Spracherkennungsprozesses durch Maximierung der Viterbi-Pfadwahrscheinlichkeiten neu bestimmt.

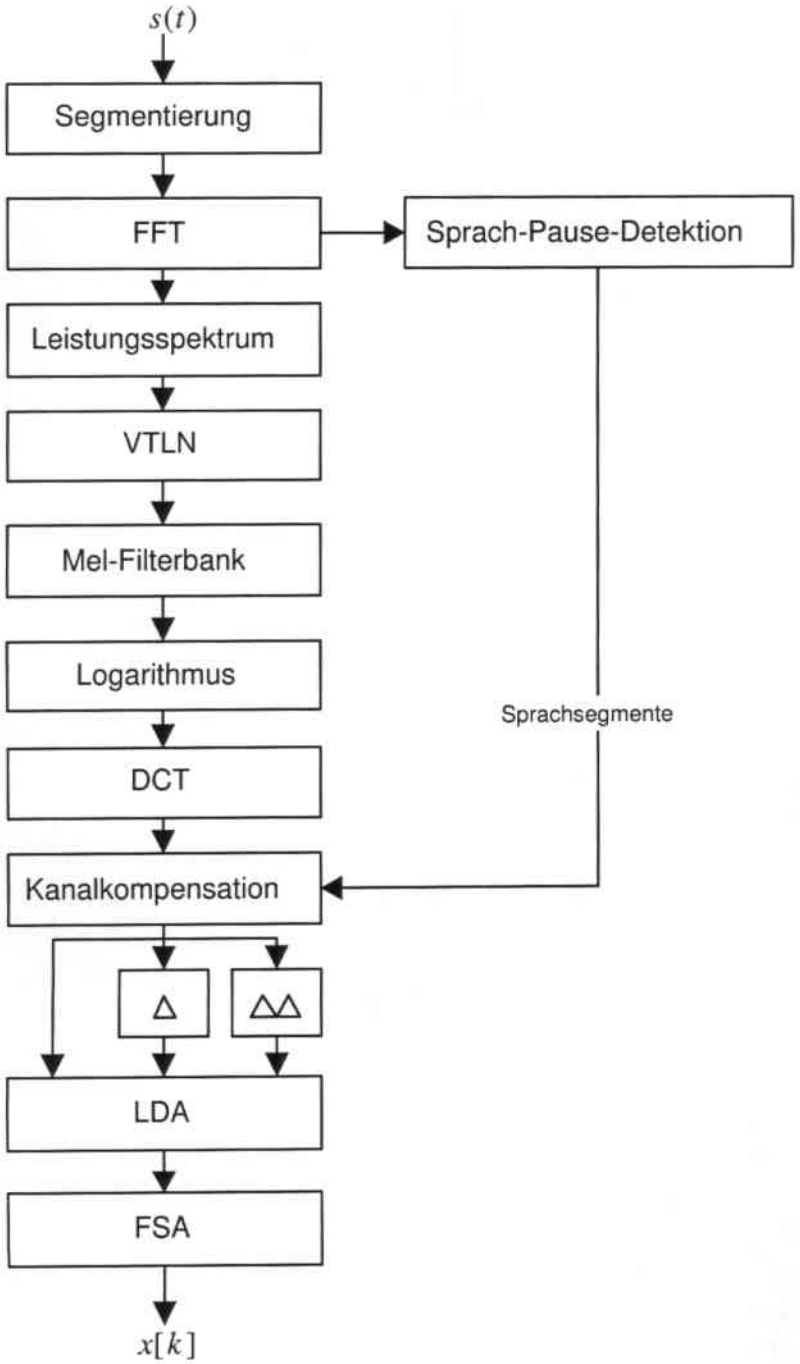


Abbildung 3.1: Stufen der Standard-Vorverarbeitung

6. Mel-Filterbank
   Das Leistungsspektrum wird nach der Vokaltraktlängennormierung durch eine Mel-Filterbank zu 30 Bändern zusammengefasst. Diese Zusammenfassung von Frequenzbändern ist der Arbeitsweise des menschlichen Gehörs nachemfpunden.

7. Logarithmierung und diskrete Cosinus-Transformation (DCT)
   Nach der Logarithmierung wird durch die diskrete Cosinus-Transformation das Cepstrum berechnet. Die so erhaltenen Vektoren werden durch das sogenannte „Liftering" auf 13 Koeffizienten beschränkt.

8. Kanalkompensation
   Die Kanalkompensation wird mit Hilfe der sprachbasierten cepstralen Mittelwertsubtraktion (SCMS) durchgeführt. Diese wird in Abschnitt 3.5 beschrieben.

9. Ableitung
   Der nach der Kanalkompensation erhaltene cepstrale Merkmalsvektor wird durch dessen erste und zweite Ableitung erweitert, da zeitliche Veränderungen der cepstralen Spektren im Erkennungsprozess von Bedeutung sind.

10. Lineare Diskriminanzanalyse
    Mit Hilfe der Linearen Diskriminanzanalyse (LDA) wird der Merkmalsvektor in seiner Dimension – und damit auch die enthaltene Redundanz – reduziert. Durch dieses Verfahren werden die Vektoren weiterhin in eine für den Spracherkennungsprozess vorteilhaftere Darstellung transformiert. Hierzu wird beim Training der Akustik eine LDA-Matrix erstellt. Die Merkmalsvektoren werden beim Training zu Lautklassen zusammengefasst, wobei die durchschnittliche Varianz innerhalb einer Klasse minimiert und die Varianz zwischen den Klassen maximiert wird. Dadurch erhöht sich die Trennschärfe zwischen den Klassen, was die Klassifikation der extrahierten Merkmalsvektoren vereinfacht. Mit Hilfe dieser LDA-Matrix wird in diesem Vorverarbeitungsschritt eine Hauptachsentransformation des Merkmalsvektors durchgeführt.

11. FSA
    Im letzten Schritt kann eine Anpassung an das Signal mittels FSA[2] durchgeführt werden. Um die Anpassung berechnen zu können, wird zuerst eine vorgegebene Anzahl an Merkmalsvektoren über eventuell mehrere Äußerungen gesammelt. Ist die vorgegebene Anzahl an Merkmalsvektoren erreicht, wird bei jeder neu verarbeiteten Äußerung eine Transformationsmatrix berechnet. Dafür wird über alle Modelle der Viterbi-Pfad der gesammelten Merkmalsvektoren bestimmt, um festzustellen, welche phonetischen Modelle verwendet werden. Durch eine weitere Wahrscheinlichkeitsberechnung pro Modell wird eine n-besten-Liste der Referenzvektoren jedes Codebuchs aufgestellt. Von dieser werden die n-besten Referenzvektoren sowie die zugehörigen Varianzen ausgewählt, um, wie in [Gale98] angegeben, eine Transformationsmatrix durch eine Iteration über alle Zeilen der Matrix unter Maximierung der Likelihood zu berechnen. Dies ist notwendig, da die einzelnen Zeilen der Transformationsmatrix voneinander abhängig sind. Die berechnete Transformationsmatrix wird in den

[2]Feature Space Adaptation

Merkmalsraum transformiert und dort auf die Merkmalsvektoren der folgenden Äußerungen angewandt.

Bevor die einzelnen Geräuschreduktionsverfahren vorgestellt werden, folgt eine Darstellung der Sprachaufnahmesituation mit Hilfe des Kanalmodells.

3.2 Kanalmodell

Das Modell in Abbildung 3.2 beschreibt die Ausgangssituation der aufgenommenen Sprache. Die Akustik des Raums, das Mikrofon des Headsets und die zur Aufnahme der Sprache benutzte Hardware bestimmen die Übertragungseigenschaften $\tilde{h}$ des Kanals. Unter der Annahme, dass diese Übertragungseigenschaften während einer Aufnahme linear und zeitinvariant sind, läßt sich das Signal $\tilde{y}$ im Zeitbereich als Faltung des Sprachsignals s mit der Übertragungseigenschaft $\tilde{h}$ des Kanals berechnen. Nichtlineare Effekte werden vernachlässigt.

$$\tilde{y}(t) = s(t) * \tilde{h}(t) \tag{3.1}$$

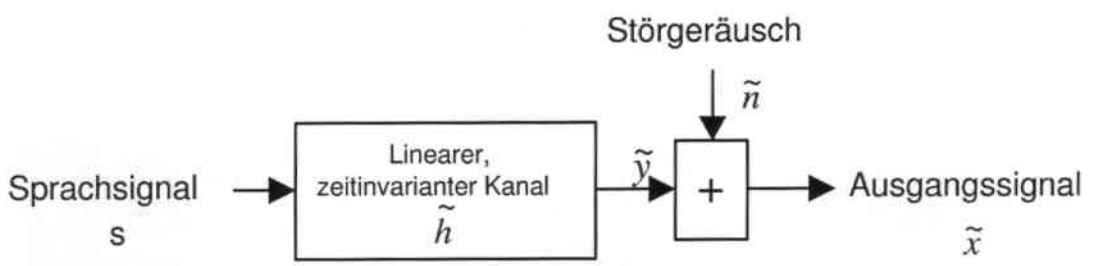


Abbildung 3.2: Kanalmodell

Hintergrund- beziehungsweise Störgeräusche sowie Atemgeräusche bei Nahbesprechungsmikrofonen überlagern additiv das Signal $\tilde{y}$. Somit kann die Gleichung 3.1 um die Störgeräusche $\tilde{n}$ erweitert werden. Solange die Störgeräusche keine Sprache beinhalten, sind diese nicht mit dem Sprachsignal korreliert. Das Ausgangssignal $\tilde{x}$ berechnet sich unter Einbeziehung der Störgeräusche und des Übertragungskanals als

$$\tilde{x}(t) = s(t) * \tilde{h}(t) + \tilde{n}(t) \tag{3.2}$$

Berechnet man aus Gleichung 3.2 die Fourier-Transformierte so ändert sich bei diesem Übergang in den Spektralbereich die Gleichung wie folgt:

$$\widetilde{X}(\omega) = S(\omega) \cdot \widetilde{H}(\omega) + \widetilde{N}(\omega) \tag{3.3}$$

Bei einem weiteren Übergang in das Betragsleistungsspektrum bleibt dieser Zusammenhang für die einzelnen Koeffizienten i des Vektors für die segmentierten Sprachrahmen k erhalten.

$$\widetilde{X}_i[k] = S_i[k] \cdot \widetilde{H}_i[k] + \widetilde{N}_i[k] \tag{3.4}$$

mit $i = 1 \ldots N$, Anzahl der Koeffizienten eines Vektors.

Hier wird deutlich, dass durch Verfahren, wie beispielsweise die spektrale Subtraktion, im Spektralbereich das Störgeräusch nach einer entsprechenden Geräuschschätzung vom Ausgangssignal subtrahiert und somit idealerweise entfernt werden kann. Nach dieser Geräuscheliminierung beziehungsweise unter Vernachlässigung der Störgeräusche und dem Übergang zu logarithmierten Mel-Koeffizienten erhält man vereinfacht den Zusammenhang

$$\log\left(\widetilde{X}_i[k]\right) = \log\left(S_i[k]\right) + \log\left(\widetilde{H}_i[k]\right) \tag{3.5}$$

Dieser Zusammenhang bleibt auch beim Übergang in den Cepstralbereich erhalten. Somit kann durch eine weitere Subtraktion, der cepstralen Mittelwertsubtraktion, beschrieben in Abschnitt 3.5, der Kanal kompensiert werden.

Westphal [West01] erweiterte das Kanalmodell zum Umgebungsmodell, in dem er zusätzlich zum Übertragungskanal den Sprecherkanal und die unterschiedlichen Vokaltraktlängen der Sprecher berücksichtigte. In dieser Arbeit wird jedoch die Vokaltraktlänge nicht berücksichtigt, da die Vokaltraktlängennormierung zur Standard-Vorverarbeitung gehört. Der Sprecherkanal wird nicht explizit aufgeführt. Er wird den Eigenschaften des Übertragungskanals zugeordnet.

Weiterhin können durch hohe Geräuschanteile kleine Werte des Sprachsignals überdeckt und somit ihr Informationsgehalt vernichtet werden. Dies führt zu fehlenden Merkmalen des Sprachsignals (Missing Features), was sich negativ auf die Erkennungsleistung des Spracherkennungssystems auswirkt. Dieser Umstand wurde in dieser Arbeit ebenfalls nicht berücksichtigt.

3.3 Spektrale Subtraktion

Die spektrale Subtraktion ist ein bekanntes und häufig genutztes Verfahren zur Minderung der Einflüße additiver Störgeräusche. In [Boll79] wurde dieses Verfahren bereits 1979 zur Geräuschreduktion vorgeschlagen. Voraussetzung bei diesem Verfahren ist, dass das Störgeräusch das Nutzsignal nur additiv überlagert und nicht mit diesem korreliert ist. In dieser Arbeit wurde die spektrale Subtraktion als einkanaliges Verfahren mit Geräuschschätzung und als zweikanaliges Verfahren betrachtet.

Einkanalige spektrale Subtraktion

Wie bereits im vorhergehenden Abschnitt in den Gleichungen 3.3 und 3.4 gezeigt wurde, läßt sich das additive Störgeräusch durch eine Subtraktion eines geschätzten Störgeräusches $\hat{n}$ im Spektralbereich verringern. Bei dieser Schätzung wird angenommen, dass das Störgeräusch quasi stationär ist. Damit wird auch eine geringfügige Varianz des Störgeräuschs angenommen. Unter diesen Annahmen läßt sich das Störgeräusch durch eine Mittelung der gestörten Merkmalsvektoren über alle Sprachpausen berechnen. Dieser Mittelwert ist somit unabhängig von der Zeit und wird mit Hilfe der spektralen Subtraktion von allen gestörten Merkmalsvektoren subtrahiert.

Eine weiteres Geräuschschätzungsverfahren ist in [Mart93] beschrieben. Die Störgeräuschschätzung erfolgt über das Verfahren der Minimum Statistik. Dabei wird über ein Zeitfenster und nochmaliger Unterteilung dieses Fensters das jeweilige Minimum der geglätteten Koeffizienten der Energie pro Frequenzband bestimmt. Der Schätzwert des Störgeräuschs ergibt sich dann durch Multiplikation eines Überschätzfaktors mit diesem Minimum. Voraussetzung für dieses Verfahren ist die statistische Unabhängigkeit des Sprachsignals vom Störgeräuschsignal. Es berücksichtigt schnell und langsam verändernde Störsignale und benötigt keinen Sprach-Pause-Detektor. Dieses Verfahren zeigte in [West01] gute Ergebnisse und wurde deshalb in dieser Arbeit übernommen. Auf die Störgeräuschschätzung über die Mittelung der gestörten Merkmalsvektoren wurde verzichtet.

Nach der Geräuschschätzung kann ein „entstörtes" Signalspektrum $\tilde{y}[k]$ aus dem gestörten Signalspektrum $x[k]$ und dem geschätzten Spektrum des Störgeräuschs $\hat{n}[k]$ berechnet werden. Um negative Werte zu verhindern wird ein sogenannter Spectral Floor b definiert, der mit dem Signalspektrum $x[k]$ multipliziert wird. Das Ergebnis dieser Multiplikation dient bei negativen Werten der Subtraktion als Ersatzwert. a ist als Overestimation Factor definiert. Mit diesem Wert wird die „Subtraktions-Stärke" des Störgeräuschspektrums gesteuert.

$$\tilde{y}[k] = \begin{cases} x[k] - a \cdot \hat{n}[k], & x[k] - a \cdot \hat{n}[k] \geq b \cdot x[k] \\ b \cdot x[k], & x[k] - a \cdot \hat{n}[k] < b \cdot x[k] \end{cases} \tag{3.6}$$

Diese Methode besitzt allerdings einen Nachteil. Da das Störgeräuschspektrum geschätzt wird (Mittelung bei der Geräuschschätzung), können im Signalspektrum nach der spektralen Subtraktion Reste des Rauschens in einigen Spektralbändern übrig bleiben. Diese werden als „Musical Tones" bezeichnet.

Die Verwendung dieser einkanaligen Geräuschreduktionsmethoden bietet sich vor allem bei Anwendungen an, bei denen das Störgeräusch im Gegensatz zu zweikanaligen Verfahren nicht mit Hilfe eines zweiten Mikrofons isoliert aufgenommen werden kann. Dies wäre zum Beispiel bei einem per Sprache über Funk gesteuerten Flugplatz oder dem automatischen Abhören der ATIS[3] durch das FMS[4] – beides noch Visionen des Autors – der Fall. Ein weiterer Nachteil des einkanaligen Verfahrens liegt in der statistischen Schätzung des Störgeräuschs, welches ein ungenaues Modell des Störgeräuschs zur Folge hat.

Zweikanalige spektrale Subtraktion

Während bei der einkanaligen Version der spektralen Subtraktion das Störgeräusch geschätzt werden muss, was fehlerbehaftet ist, verspricht die zweikanalige spektrale Subtraktion bessere Erkennungsergebnisse, da mit dem zweiten Aufnahmekanal reine, zeitlich passende Störgeräuschaufnahmen vorliegen. Allerdings ist hierzu im Gegensatz zur einkanaligen Version ein erhöhter Hardware-Aufwand zur Aufnahme zu betreiben.

Das Funktionsprinzip der spektralen Subtraktion in der zweikanaligen Version ist das selbe wie in der einkanaligen Version. Nur wird in der zweikanaligen Version nicht das

[3] Automatische Ausstrahlung von Start- und Landeinformationen an Flugplätzen
[4] Flight Management System

geschätzte Störgeräusch, sondern das real parallel aufgezeichnete Geräusch nach ein paar Anpassungsoperationen subtrahiert. Voraussetzung ist natürlich die parallele (stereo) Aufzeichnung der gestörten Sprache und des Störgeräuschs.

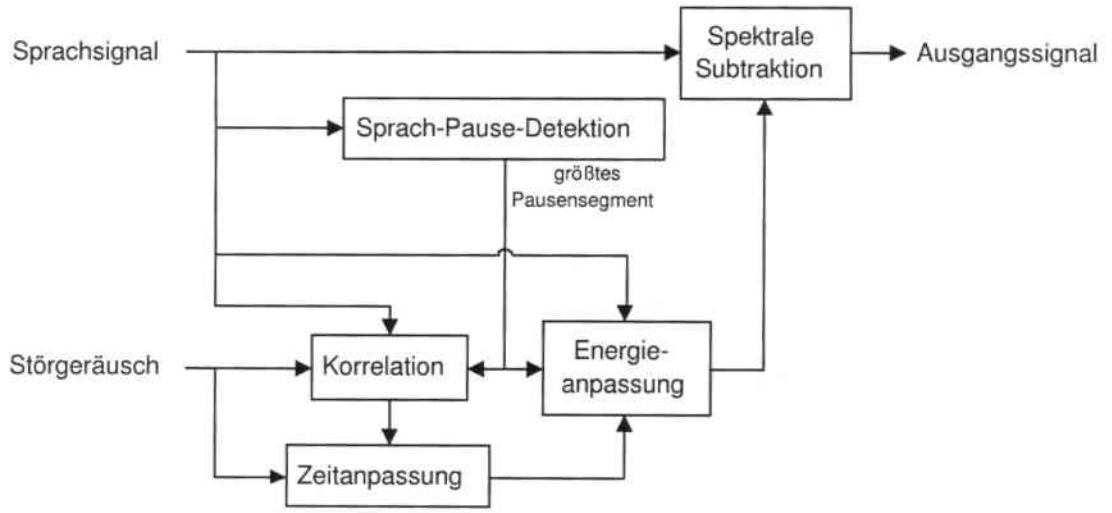


Abbildung 3.3: Spektrale Subtraktion mit zwei zur Verfügung stehenden Aufnahmekanälen

Abbildung 3.3 zeigt die Vorgehensweise bei der zweikanaligen spektralen Subtraktion. Unter Verwendung der in Abschnitt 3.6 vorgestellten bereichskombinierten Sprach-Pause-Detektion werden aus der gestörten Sprachaufnahme die Grenzen des größten Pausensegments festgestellt. Innerhalb des so festgestellten Anfangs- beziehungsweise Endpunktes erfolgt zunächst eine Anpassung der reinen Geräuschaufnahme an die Sprachaufnahme. Das Pausensegment des Sprachsignals p_{speech} wird mit dem korrespondierenden Signalsegment der Geräuschaufnahme p_{noise} korreliert. Durch die Bestimmung des Maximums der Korrelationsfunktion wird der zeitliche Verschiebungsfaktor f_t festgestellt.

$$f_t = \max_{\text{alle sinnvollen Segmentverschiebungen}} \left\{ \sum_{k=0}^{N-1} p_{speech}[k] \cdot p_{noise}[N-k] \right\} \tag{3.7}$$

mit der Länge N des aus dem Signal extrahierten Pausensegments.

Bei der Verschiebung des Geräuschsegments wird auch das Pausensegment des Sprachsignals beschnitten, so dass eine Verfälschung der Korrelationsergebnisse durch die sonst notwendige Einführung von Nullvektoren vermieden wird.

Mit diesem Verschiebungsfaktor f_t wird danach die Geräuschaufnahme verschoben und wiederum das korrespondierende Segment bestimmt. Innerhalb dieser Segmente erfolgt nun die Ermittlung eines Anpassungsfaktors f_e im Spektralbereich zur Anpassung der Energie des Geräuschsegments an die Energie des Pausensegments der Sprachaufnahme mit Hilfe der Methode der kleinsten Fehlerquadrate.

$$f_e = \min_{\alpha \in [-1,0...1,0]} \left\{ \sum_{k=0}^{N-1} \left(P_{speech}[k] - \alpha \cdot P_{noise}[k] \right)^2 \right\} \tag{3.8}$$

mit der Anzahl N der Spektralvektoren des Pausensegments.

Alle Vektoren der Geräuschaufnahme werden dann mit dem Anpassungsfaktor f_e multipliziert. Danach erfolgt die spektrale Subtraktion von Sprachaufnahme und angepasster Geräuschaufnahme, wie es bereits im vorangegangenen Abschnitt beschrieben wurde.

Zahlreiche weitere Verfahren sind möglich, unter anderem auch die in Abbildung 3.4 gezeigte Filterung mit einem adaptivem Filter wie es in [IlKa99] oder [Sing01] beschrieben wird. Diese wurden jedoch nicht in diese Arbeit einbezogen, da sich im Verlauf der Experimente die gemachten zweikanaligen Aufnahmen als unbrauchbar herausstellten.

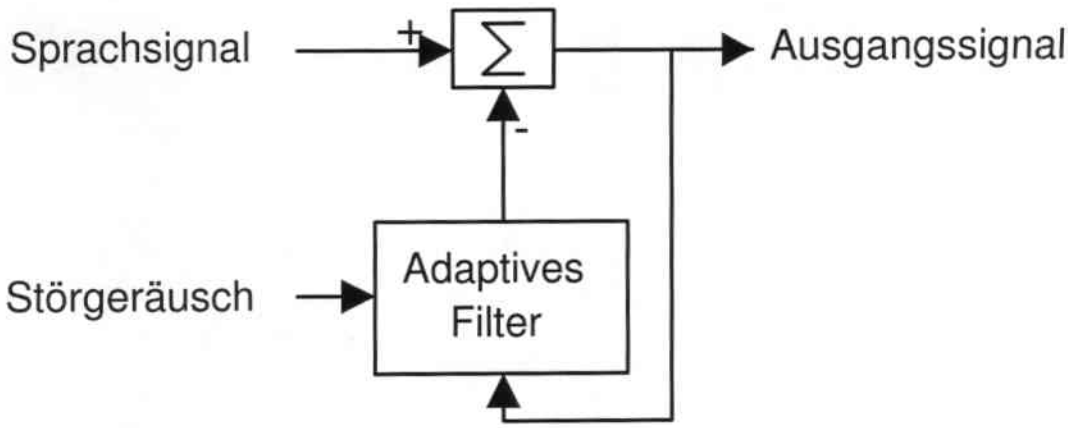


Abbildung 3.4: Zweikanalige Geräuschreduktion mit einem adaptiven Filter

3.4 MAM — Modellkombinationsbasierte akustische Transformation

Die modellkombinationsbasierte akustische Transformation (MAM[5]) ist ein von Westphal in [West01] vorgestelltes Verfahren zur Kompensation der Umgebungseinflüsse. Primär wurde es für den Einsatz im Auto bei Aufnahmen mit einem Fernbesprechungsmikrofon entwickelt. Es wurde jedoch auch für andere Geräuschumgebungen vorgeschlagen. Entwicklungsziel war eine robuste Spracherkennung in einer sich ständig ändernden Umgebung. Diese sich ständig verändernde Umgebung ist nur durch die Störgeräusche der aktuell bearbeiteten Aufnahme charakterisiert. Deshalb erfolgt die Kompensation individuell für jede Äußerung.

Voraussetzung für das Verfahren ist die Gültigkeit des in Abschnitt 3.2 angesprochene Umgebungsmodells: Der Kanal bleibt für die Dauer der Aufnahme konstant. Abweichend von Westphal wird hier als Voraussetzung nur das vorgestellte Kanalmodell festgelegt, da eine Vokaltraktlängennormierung, wie sie Westphal in dem Verfahren verwendet hat, zur Standard-Vorverarbeitung gehört und somit nicht mehr als Teil des Verfahrens betrachtet wird. Auch wird die Bestimmung des Verzerrungsfaktors nicht betrachtet, da diese ebenfalls zum Standard der Spracherkennung zu zählen ist.

[5]Model Combination Based Acoustic Mapping

Weitere Voraussetzungen sind kaum verrauschte Trainingsaufnahmen des Spracherkenners, eine Normierung von Kanal und Sprecher während des Trainingsprozesses und die Verwendung eines akustischen Modells im Spracherkenner, was auf einer homogenen Sprachdatenbasis trainiert wurde. Aus den Trainingsdaten muss vor einer Anwendung des Verfahrens noch ein Modell ungestörter Sprache λ trainiert werden, was nur eine phonetische Klasse besitzt und im Log-Spektralbereich trainiert wird.

$$\lambda = \{\mu, C\} \tag{3.9}$$

Eine Übersicht über das Verfahren gibt Abbildung 3.5. Anhand dieser werden nun die einzelnen Schritte erläutert.

Schritt 1: Sprach-Pause-Detektion

Zuerst werden mit Hilfe eines Sprach-Pause-Detektors die Sprach- und Pausensegmente des Sprachsignals festgestellt. Dies ist für die im nächsten Schritt durchgeführte Kanalkompensation notwendig. Weiterhin werden für Schritt 3 die Pausensegmente benötigt. Die weiteren Schritte werden nur angewandt, falls mehr als 10 Störgeräuschrahmen in der betrachteten Äußerung vorhanden sind. Dies ist notwendig, da bei weniger als 10 Störgeräuschrahmen das Geräuschmodell in Schritt 3 nicht ausreichend geschätzt werden kann.

Schritt 2: Kanalkompensation im Log-Spektralbereich

Im Log-Spektralbereich wird die Kanalkompensation durch eine sprachbasierte Mittelwertsubtraktion, wie sie im nachfolgenden Abschnitt für den cepstralen Bereich beschrieben wird, durchgeführt. Dies ist aufgrund der Betrachtungen im Abschnitt 3.2 (Gleichung 3.5) möglich. Die Kanalkompensation wird vor der Geräuschkompensation durch die modellbasierte akustische Transformation durchgeführt, um den Modellierungsaufwand für das im nächsten Schritt verwendete Modell zu verringern und die Schätzung des Geräuschmodells zu verbessern.

Schritt 3: Geräuschmodellschätzung

Die durch den Sprach-Pause-Detektor festgelegten Pausensegmente werden im Log-Spektralbereich ausgeschnitten und zusammengefasst. In der Sprach-Pause-Detektion muss dabei eine Fehlklassifikation von Sprachsegmenten als Pausensegmente ausgeschlossen werden. Damit steht das reine Störgeräusch zur Verfügung. Das Störgeräusch kann nun durch eine Gaußdichte mit dem Mittelwert $\tilde{\mu}$ und der Kovarianzmatrix $\tilde{C}$ modelliert werden. Dieses Modell $\tilde{\lambda}$ der Störgeräusche wird im Gegensatz zum Modell λ ungestörter Sprache nur auf der aktuellen bearbeiteten Sprachaufnahme bestimmt. Die Modellbestimmung kann nur durchgeführt werden, wenn in der bearbeiteten Äußerung mehr als 10 Pausensegmente vorhanden sind. Sind weniger als 10 Pausensegmente vorhanden, kann der Mittelwert und die Varianz des Geräuschmodells nicht mehr adäquat bestimmt werden.

$$\tilde{\lambda} = \left\{\tilde{\mu}, \tilde{C}\right\} \tag{3.10}$$

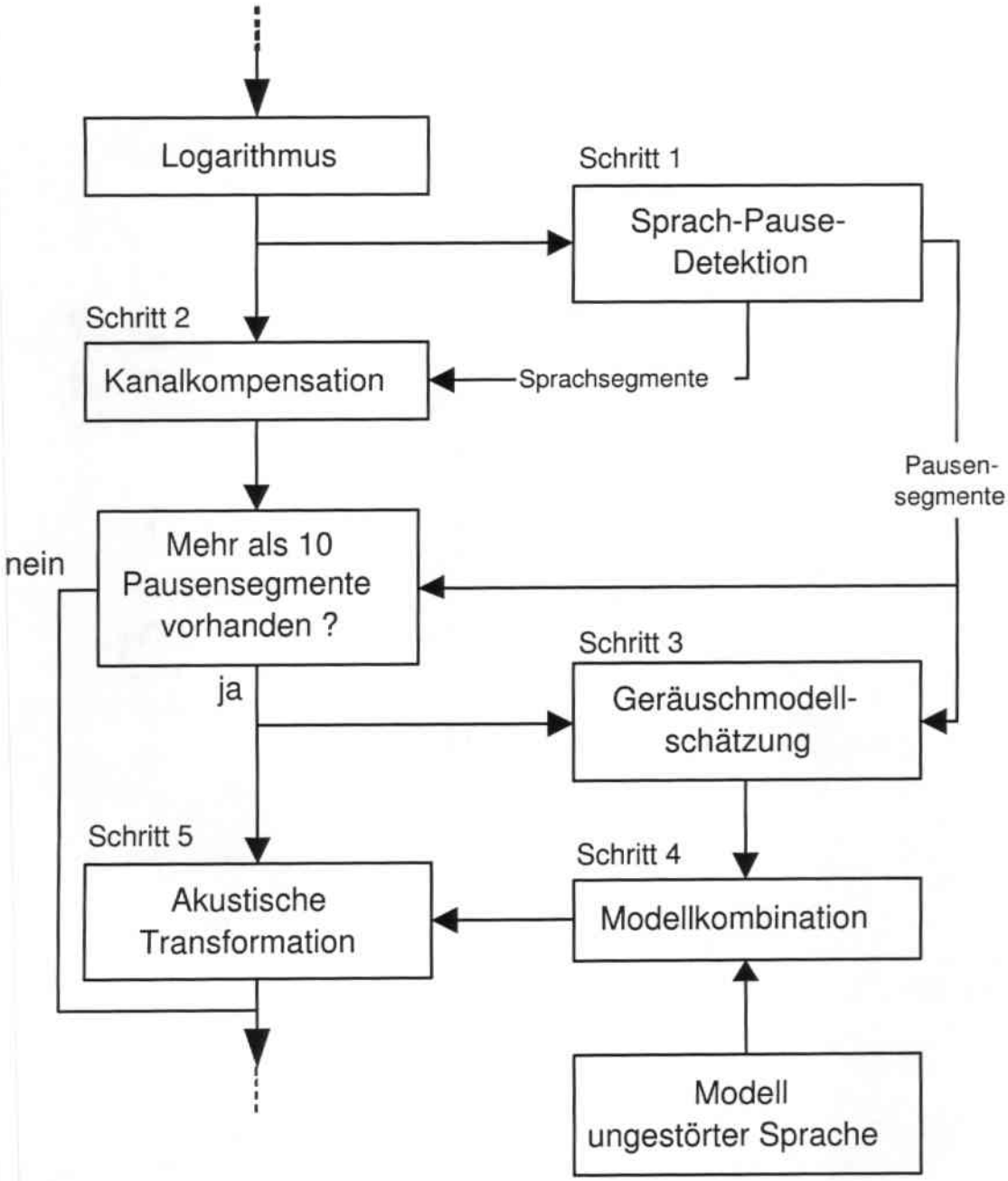


Abbildung 3.5: Modellkombinationsbasierte Akustische Transformation

Schritt 4: Modellkombination

In diesem Schritt wird das Modell λ ungestörter Sprache mit dem im vorhergehenden Schritt trainierten Modell $\tilde{\lambda}$ der Störgeräusche mit Hilfe der Parallel Model Combination (PMC) kombiniert.

$$\tilde{\lambda}_{PMC} = PMC\left(\lambda, \tilde{\lambda}\right) \tag{3.11}$$

Durch die Kombination zweier akustischer Modelle erhält man durch die PMC wieder ein herkömmliches akustisches Modell. Die Topologie des Sprachmodells ändert sich dabei bei einem Geräuschmodell mit einer Gaußdichte nicht. Die PMC läuft folgendermaßen ab:

1. Transformation in den linearen Spektralbereich
2. Kombination der Modelle
   Bei nur einer Normalverteilung für das Geräuschmodell vereinfacht sich die Kombination auf die Addition der Mittelwerte und Kovarianzmatrizen.

$$\mu_{PMC} = \mu + \tilde{\mu} \tag{3.12}$$
$$C_{PMC} = C + \tilde{C} \tag{3.13}$$
$$\tilde{\lambda}_{PMC} = \{\mu_{PMC}, C_{PMC}\} \tag{3.14}$$

   Besteht das Geräuschmodell aus mehr als einer Normalverteilung, so wird ein datengetriebener Ansatz verfolgt. Hierzu werden aus den beiden Modellen künstlich Beispiele generiert. Diese Beispiele werden dann addiert und zur Schätzung des kombinierten Modells verwendet.
3. Transformation in den Log-Spektralbereich

Schritt 5: Akustische Transformation

Mit Hilfe des im vorhergehenden Schritt kombinierten Modells $\tilde{\lambda}_{PMC}$ wird nun durch die Akustische Transformation aus jedem Merkmalsvektor $\tilde{x}[k]$ des Analyserahmen k der Äußerung ein dem korrespondierenden Modell λ der ungestörten Sprache entsprechender Merkmalsvektor $\hat{s}[k]$ geschätzt.

$$\hat{s}[k] = AM\left(\tilde{x}[k], \tilde{\lambda}_{PMC}, \lambda\right) \tag{3.15}$$

Die durch die Sprachaufnahme und die Vorverarbeitung erhaltenen gestörten Merkmalsvektoren $\tilde{x}[k]$ lassen sich durch eine Verschiebungsoperation aus ihren korrespondierenden ungestörten Merkmalsvektoren $x[k]$ berechnen. Dabei hängt der Verschiebungsvektor Δ von dem ungestörten Merkmalsvektor $x[k]$ ab.

$$\tilde{x}[k] = x[k] + \Delta\left(x[k]\right) \tag{3.16}$$

Der ungestörte Merkmalsvektor $x[k]$ liegt jedoch nicht vor. Statt dessen soll er durch das Verfahren angenähert werden. Durch die vorausgehenden Schritte stehen zwei Modelle zur Verfügung: eines für ungestörte (λ) und eines für

gestörte Sprache $\left(\tilde{\lambda}_{PMC}\right)$. Die Mittelwerte ($\mu$) dieser Modelle stellen prototypische Vektoren dar, aus denen Verschiebungsvektoren berechnet werden können. Diese Verschiebungsvektoren werden mit der bedingten Klassenwahrscheinlichkeit gewichtet und bilden in der Summe den gesuchten Korrekturvektor. Die Klassenwahrscheinlichkeit kann über die Auswertung einer Gaußdichte des Modells für gestörte Sprache berechnet werden.

Diese Methode der Schätzung des ungestörten Merkmalsvektors mit Hilfe des Korrekturvektors bezeichnet Westphal als „Akustische Transformation".

Eine Schwachstelle des eine verändernde Umgebung und damit veränderte Störgeräusche kompensierenden Verfahrens ist die Sprach-Pause-Detektion. Westphal benutzte dazu einen einfachen energiebasierten Sprach-Pause-Detektor. Dieser erkannte durch einen vorher fest eingestellten Schwellwert die Sprachpausen, die dann für das Training des Störgeräuschmodells $\tilde{\lambda}$ verwendet wurden. Dieser Sprach-Pause-Detektor muss daher für sich gravierend ändernde Störgeräusche, wie sie im Luftfahrzeug auftreten können, neu eingestellt werden. In Versuchen erwies sich der verwendete energiebasierte Sprach-Pause-Detektor als nicht praktikabel, da dieses Verfahren auf den Luftfahrzeugdaten eine ständige Anpassung des konstanten Schwellwertes verlangte. Dies ist nicht adaptiv, so dass der Sprach-Pause-Detektor gegen den in Abschnitt 3.6 vorgestellten bereichskombinierten Sprach-Pause-Detektor ausgetauscht wurde. Eine Kanalkompensation durch die sprachbasierte cepstrale Mittelwertsubtraktion wurde aufgrund von Verbesserungen bei Versuchen ebenfalls eingeführt. Diese Verbesserung kann auf die Veränderung der Merkmalsvektoren durch die akustische Transformation zurückgeführt werden. Einen Überblick über diese Modifikation gibt die Abbildung 4.5 in Kapitel 4.6.

Nachdem im Abschnitt 3.3 die zweikanalige spektrale Subtraktion betrachtet wurde, soll hier nicht unerwähnt bleiben, dass eine Verwendung eines reinen Geräuschkanals sich auch bei der MAM anbietet. Durch dieses Vorgehen ist aber trotzdem die Verwendung eines Sprach-Pause-Detektor notwendig. Das bei der zweikanaligen spektralen Subtraktion vorgestellte Verfahren kann auch hier angewendet werden. Die in Gleichung 3.7 betrachete Korrelation muss durchgeführt werden, um danach mit Gleichung 3.8 einen Anpassungsfaktor zu berechnen. Nach der Anpassung und Transformation des reinen Geräuschkanals in den Log-Spektralbereich und Mittelwertsubtraktion kann die Schätzung beziehungsweise das Training des Störgeräuschmodells durchgeführt werden. Alle weiteren Schritte folgen dann ab Schritt 4 wie oben beschrieben.

3.5 Cepstrale Mittelwertsubtraktion

Die cepstrale Mittelwertsubtraktion wird zur Kanalkompensation eingesetzt. In Abschnitt 3.2 wurde in Gleichung 3.5 der additive Zusammenhang zwischen der Sprache und den Kanaleigenschaften dargestellt. Dieser Zusammenhang bleibt beim Übergang in den cepstralen Bereich erhalten. Die Gleichung kann im cepstralen Bereich wie folgt geschrieben werden.

$$\tilde{x}[k] = s[k] + \tilde{h}_{Kanal} \tag{3.17}$$

mit der Bezeichnung k für einen Analyserahmen.

Um die cepstrale Mittelwertsubtraktion durchführen zu können, muss zuerst der Mittelwert der Analyserahmen berechnet werden. Nach [West01] bleibt der Kanalbeitrag dabei erhalten, da der Kanal als zeitinvariant angenommen wird.

$$\frac{1}{N} \cdot \sum_{k=0}^{N-1} \tilde{x}[k] = \tilde{h}_{Kanal} + \frac{1}{N} \cdot \sum_{k=0}^{N-1} s[k] \tag{3.18}$$

Die rechte Summe dieser Gleichung ist nun hauptsächlich vom Sprecher und der Lautverteilung abhängig. Nimmt man weiterhin an, dass die Verteilung der Laute und der Sprecher konstant sind, so kann man diese Summe als „Sprecherkanal" $\tilde{h}_{Sprecher}$ bezeichnen und diesen mit dem eigentlichen Kanal zu einem kombinierten Kanal $\tilde{h}$ zusammenfassen.

$$\frac{1}{N} \cdot \sum_{k=0}^{N-1} \tilde{x}[k] = \tilde{h}_{Kanal} + \tilde{h}_{Sprecher} = \tilde{h} \tag{3.19}$$

Wird nun dieser so erhaltene Kanal $\tilde{h}$ von dem gestörten Ausgangssignal $\tilde{x}[k]$ im cepstralen Bereich subtrahiert, so bleibt der mittelwertfreie Sprachanteil $\hat{s}[k]$ übrig. Dieser Teil trägt die für den Erkennungsprozeß wichtigen Lautinformationen. Dieses Verfahren wird cepstrale Mittelwertsubtraktion genannt.

$$\text{CMS: } \hat{s}[k] = \tilde{x}[k] - \frac{1}{N} \cdot \sum_{l=0}^{N-1} \tilde{x}[l] = \tilde{x}[k] - \bar{x} \tag{3.20}$$

Ein Nachteil der cepstralen Mittelwertsubtraktion ist die starke Abhängigkeit von dem Störgeräusch und der Dauer der Sprachpausen. Aus diesem Grund wurde die einfache cepstrale Mittelwertsubtraktion in dieser Arbeit nicht verwendet. Um die Nachteile zu kompensieren, wird die Mittelwertbildung auf die Sprachrahmen beschränkt. Diese sprachbasierte cepstrale Mittelwertsubtraktion (SCMS) wird in der Standard-Vorverarbeitung verwendet. Allerdings ist zur Mittelwertbildung über die Sprachrahmen ein Sprach-Pause-Detektor notwendig. Dieser liefert eine binäre Entscheidung $w[k]$: 1 für Sprachrahmen und 0 für Pauserahmen.

$$\text{SCMS: } \hat{s}[k] = \tilde{x}[k] - \frac{\sum_{l=0}^{N-1} w[l] \cdot \tilde{x}[l]}{\sum_{l=0}^{N-1} w[l]} = \tilde{x}[k] - \bar{x}^{Sprache} \tag{3.21}$$

Da bei der SCMS Sprachpausen nicht berücksichtigt werden, kann man zusätzlich einen Mittelwert über die Sprachpausen bilden und diesen nur von den Sprachpausen abziehen, um die Einflüsse der Störgeräusche zu kompensieren. Dies ist dann die 2-Level cepstrale Mittelwert-Subtraktion (2CMS).

$$\begin{aligned} \text{2CMS: } \hat{s}[k] &= \tilde{x}[k] - w[k] \cdot \frac{\sum_{l=0}^{N-1} w[l] \cdot \tilde{x}[l]}{\sum_{l=0}^{N-1} w[l]} \\ &\quad - (1 - w[k]) \cdot \frac{\sum_{l=0}^{N-1} (1 - w[l]) \cdot \tilde{x}[l]}{\sum_{l=0}^{N-1} (1 - w[l])} \\ &= \tilde{x}[k] - w[k] \cdot \bar{x}^{Sprache} - (1 - w[k]) \cdot \bar{x}^{Pause} \end{aligned} \tag{3.22}$$

Durch die Subtraktion des Mittelwertes werden nun die Sprachsegmente wie auch die Pausensegmente nach [West01] auf den gleichen Punkt verschoben, wo sie sich überlagern. Steht kein weiteres Maß zur Verfügung, so können sich dadurch die Einfügefehler bei dem Spracherkennungsprozeß durch eine Verwechslung von Pausen mit Sprachsegmenten erhöhen. Deshalb wurde auch dieses Verfahren in dieser Arbeit nicht benutzt.

Der Ansatz der 2CMS kann erweitert werden, in dem die Subtraktion durch eine Normierung ersetzt wird und anstatt der binären Ausgabe des Sprach-Pause-Detektors eine kontinuierliche Gewichtungsfunktion $w[k]$ verwendet wird. Die Mittelwerte der Sprach- und Pausensegmente werden nun nicht mehr zu einem gemeinsamen Punkt verschoben, sondern zum Durchschnitt der Sprach- und Pausensegmente der Akustik-Trainingsdaten. Die Differenz (Delta Δ) des aktuellen Mittelwerts der zu erkennenden Äußerung und des durchschnittlichen Mittelwerts der Trainingsmenge wird als Ausgleichsvektor verwendet. Dieses Verfahren wird zweifache Deltacepstrale Mittelwertsubtraktion (2DCMS) genannt.

$$\mu_{Sprache} = \frac{1}{M} \cdot \sum_{u=1}^{M} \bar{x}_u^{Sprache} \tag{3.23}$$

$$\mu_{Pause} = \frac{1}{M} \cdot \sum_{u=1}^{M} \bar{x}_u^{Pause} \tag{3.24}$$

$$\begin{aligned} \text{2DCMS: } \hat{s}[k] &= \tilde{x}[k] - w[k] \cdot \left(\frac{\sum_{l=0}^{N-1} w[l] \cdot \tilde{x}[l]}{\sum_{l=0}^{N-1} w[l]} - \mu_{Sprache} \right) \\ &\quad - (1 - w[k]) \cdot \left(\frac{\sum_{l=0}^{N-1} (1 - w[l]) \cdot \tilde{x}[l]}{\sum_{l=0}^{N-1} (1 - w[l])} - \mu_{Pause} \right) \\ &= \tilde{x}[k] - w[k] \cdot \left(\bar{x}^{Sprache} - \mu_{Sprache} \right) \\ &\quad - (1 - w[k]) \cdot \left(\bar{x}^{Pause} - \mu_{Pause} \right) \\ &= \tilde{x}[k] - w[k] \cdot \Delta\bar{x}^{Sprache} - (1 - w[k]) \cdot \Delta\bar{x}^{Pause} \end{aligned} \tag{3.25}$$

Aus den Trainingsdaten (M Äußerungen) der Akustik können durchschnittliche Mittelwerte für Sprachsegmente $\mu_{Sprache}$ und Pausensegmente μ_{Pause} durch Verwendung eines Sprach-Pause-Detektors berechnet werden. $\mu_{Sprache}$ und μ_{Pause} bilden ein sehr einfaches Modell ungestörter Sprache.

$$\lambda_{\text{ungestörte Sprache}} = \{\mu_{Sprache}, \mu_{Pause}\} \tag{3.26}$$

Dieses Modell ungestörter Sprache muss aus den Trainingsdaten vor dem Akustik-Training berechnet werden. Danach muss die Akustik für jede Gewichtungsfunktion $w[k]$ neu trainiert werden,was je nach Dauer des Trainings sehr aufwändig ist.

3.6 Sprach-Pause-Detektion

Alle betrachteten Verfahren benötigen eine Sprach-Pause-Detektion. Hierzu gibt es viele verschiedene Vorschläge. In der Standard-Vorverarbeitung wird ein einfacher Sprach-Pause-Detektor im Spektralbereich verwendet. Dieser nimmt die Unterscheidung zwischen Sprach- und Pausenrahmen aufgrund des Vergleichs der Energie des jeweiligen Analyserahmens mit einem konstanten Schwellwert vor.

Im Spektralbereich gibt es weitere Vorschläge zur Sprach-Pause-Detektion, welche im nächsten Abschnitt dargestellt werden. Danach folgt eine Betrachtung der Verfahren zur Sprach-Pause-Detektion im cepstralen Bereich. Zuletzt wird ein bereichskombinierter Sprach-Pause-Detektor vorgestellt, der im Rahmen dieser Arbeit entwickelt wurde.

Sprach-Pause-Detektion im spektralen Bereich

Sprach-Pause-Detektoren im Spektralbereich werden sehr oft in Spracherkennungssystemen eingesetzt. [KSYC+00] verwendet die gebräuchlichsten Verfahren für ein drahtloses Kommunikationssystem mit sich dynamisch verändernden Hintergrundgeräuschen. Zur Klassifikation werden die Energie des Analyserahmens k, das Verhältnis zwischen der Energie des Analyserahmens k und der Energie des mit einem FIR[6]-Filter gefilterten Analyserahmens k, der „Zero-Crossing-Rate" und die Messung der Signalspitzen herangezogen, wobei jedes Merkmal mit einem eigenen Schwellwert verglichen.

Die Energie $E[k]$ des Analyserahmens k wird berechnet als Logarithmus des ersten normalisierten Autokorrelationskoeffizienten R_0. Die Berechnung von R_0 ist in [Wölf03] oder [KaKr98] im Kapitel über Lineare Prediktion zu finden.

$$E[k] = 10 \cdot \log_{10}\left(\frac{1}{10} \cdot R_0[k]\right) \tag{3.27}$$

Die so ermittelte Energie wird mit einem adaptiven Schwellwert basierend auf dem SNR des Störgeräuschs verglichen. In [KSYC+00] ist nicht beschrieben, wie dieser adaptive Schwellwert ermittelt wird.

Das Verhältnis $R[k]$ zwischen den Energien wird durch eine Filterung des Analyserahmens k durch ein FIR-Filter mit der Grenzfrequenz 1 kHz und anschließender Division der nach der Filterung berechneten Energie $E_{filtered}[k]$ durch die vorher berechnete Energie $E[k]$ ermittelt.

$$E_{filtered}[k] = 10 \cdot \log_{10}\left(\frac{1}{N} \cdot h^T \cdot R_{Corr}\right) \tag{3.28}$$

mit der Impulsantwort h des FIR-Filters und der Töplitz-Autokorrelationsmatrix[7] R_{Corr}.

$$R[k] = \frac{E_{filtered}[k]}{E[k]} \tag{3.29}$$

[6] Finite-duration Impulse Response
[7] siehe [Wölf03] oder [KaKr98]

Dieses Verhältnis $R[k]$ wird mit zwei konstanten Schwellwerten verglichen.

Die „Zero-Crossing-Rate" $Z[n]$ wird mit Hilfe der „Vorzeichenfunktion" $sgn(x)$ auf dem mit einer Fensterfunktion multiplizierten unbearbeiteten Sprachsignal berechnet. Sie gibt die Anzahl der Nulldurchgänge des Signals x in dem Fenster (Analyserahmen) der Länge N an.

$$\begin{aligned} Z[n] &= \frac{1}{2} \cdot \sum_{k=0}^{N-1} |sgn\,(x[n-k]) - sgn\,(x[n-k-1])| \qquad (3.30) \\ \text{mit } sgn(x) &= \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases} \end{aligned}$$

Diese Anzahl $Z[n]$ wird wiederum mit zwei konstanten Schwellwerten verglichen.

Weiterhin werden in [KSYC+00] die Signalspitzen („Peakiness Measure") als weiteres Merkmal betrachtet. Diese werden, wie in [McBa95] angegeben, berechnet.

Nach der Berechnung der vorgestellten Merkmale werden Sprach- bzw. Pauserahmen durch eine boolsche Funktion aller Vergleichsergebnisse ermittelt. Diese boolsche Funktion ist das Ergebnis einer speziell gewählten Entscheidungsmatrix.

Durch die große Anzahl festgelegter und damit konstanter Schwellwerte ist die Adaptivität des Algorithmus auf eine geringe Anzahl möglicher Hintergrundgeräusche beschränkt. Der Geräuschkulisse im Luftfahrzeug hält dieser Algorithmus nicht stand. Dazu müssten alle Schwellwerte adaptiv festgelegt werden. Die hier vorgestellten Maße müssen aber nicht alle berücksichtigt werden. In [HaMa93b] wird eine Klassifikationsmatrix für Sprach- beziehungsweise Pausensegmente vorgestellt, deren Grundlage die zwei Maße „Zero-Crossing-Rate" und mittlere Energie bilden.

Neben [KSYC+00] gibt es noch viele weitere Arbeiten zur Sprach-Pause-Detektion im Spektralbereich. [ChANK01] beschreibt eine Sprach-Pause-Detektion durch eine statistische Analyse der Signale. Die umfassendste Arbeit stellt [Stad99] dar. Stadermann untersuchte neben den bereits erwähnten Methoden die Sprach-Pause-Detektion mit Hilfe von Hidden Markov Modellen (HMM) und mit Support Vector Machines, einem Verfahren des maschinellen Lernens. Es gibt natürlich noch weitere Ansätze zur Sprach-Pause-Detektion im Spektralbereich, die hier jedoch nicht mehr erläutert werden, da die Sprach-Pause-Detektion im cepstralen Bereich Vorteile gegenüber der Detektion im Spektralbereich hat.

Sprach-Pause-Detektion im cepstralen Bereich

Die Sprach-Pause-Detektion im cepstralen Bereich besitzt gegenüber der Detektion im Spektralbereich einige Vorteile. Diese sind die hohe Unabhängigkeit von Amplitude und Störgeräuschlevel [HaMa93a] sowie das Erreichen guter Ergebnisse bei niedrigen SNR-Werten und nicht-stationären Störsignalen [SoPo95]. Es wurde festgestellt, dass das Ende eines Sprachsegments durch einen einfachen Vergleich eines konstanten Schwellwertes mit einer durch ein spezielles Maß ermittelten Distanz bestimmt werden kann. Unterschiedliche Distanzmaße kann man in vielen Arbeiten finden. Diesen liegt die Annahme zugrunde, dass Störgeräuschcepstren eine kleinere Varianz besitzen als die Sprachcepstren.

In [HaMa93a] wird festgestellt, dass mit Hilfe der normierten euklidischen Distanz als Distanzmaß eine Sprach-Pause-Klassifikation im cepstralen Bereich effizienter und mit höherer Wahrscheinlichkeit durchgeführt werden kann als mit dem gleichen Distanzmaß bezüglich der Energie im Spektralbereich. Die normierte euklidische Distanz wurde wie folgt angegeben:

$$d = \frac{1}{p} \cdot \sum_{i=0}^{p-1} (c_i - \bar{c}_i)^2 \tag{3.31}$$

Dabei sind c_i und $\bar{c}_i$ zwei cepstrale Vektoren und p die Anzahl der Koeffizienten der cepstralen Vektoren. c_i kann als aktueller, zu klassifizierender Merkmalsvektor betrachtet werden. $\bar{c}_i$ ist der Mittelwert der Sprach- beziehungsweise Pausensegmente der Trainingsdaten. Diese Mittelwertvektoren müssen in einem „Training" ermittelt werden, so dass dieses Verfahren nur auf bestimmte Geräuschumgebungen angewandt werden kann. In [HaMa93b] wird ein Sprach-Endpunkt-Detektions-Verfahren wiederum mit Hilfe der euklidischen Distanz vorgestellt. Allerdings wird beim Distanzmaß auf die Normierung $1/p$ verzichtet.

$$d = \sum_{i=0}^{p-1} (c_i - \bar{c}_i)^2 \tag{3.32}$$

Sprach- beziehungsweise Pausensegmente werden durch einen Codebuch-Ansatz modelliert, welcher sich aber wiederum auf die Mittelwerte der entsprechenden cepstralen Vektoren reduzieren läßt.

In [SoPo95] werden zwei weitere Distanzmaße erwähnt. Das differentielle Maß (Gleichung 3.33) kann für die Konstruktion eines Sprach-Pause-Detektors verwendet werden. Allerdings sind weitere Berechnungen für eine brauchbare Verwendung erforderlich, die nicht angegeben wurden. Dieser ist dann nach [PoSU95] besser für die Pausendetektion als für die Sprachdetektion geeignet. Die Betrachtung des Vorgänger-Merkmalsvektors und des zu klassifizierenden Merkmalsvektors reicht nicht aus, um eine adäquate Klassifikation vornehmen zu können. Deshalb wird das differentielle Maß nicht weiter verwendet.

$$d[n] = \sqrt{\sum_{i=0}^{p-1} (c_i[n] - c_i[n-1])^2} \tag{3.33}$$

mit der Anzahl der Koeffizienten p des cepstralen Merkmalsvektors und dem Analyserahmen n.

Als weiteres Distanzmaß für die Klassifikation von Sprachrahmen wird das integrale Maß vorgestellt. Dieses berücksichtigt, dass die Energie des Analyserahmens im ersten Cepstralkoeffizienten c_0 repräsentiert ist und dominiert. Deshalb werden die verbleibenden Koeffizienten mit dem Faktor 2 gewichtet. Die Herkunft des Faktors $4,3429$ konnte weder in [SoPo95] noch in den dort angegebenen Referenzen nachvollzogen werden.

$$d = 4,3429 \cdot \sqrt{(c_0 - \bar{c}_0)^2 + 2 \cdot \sum_{i=1}^{p-1} (c_i - \bar{c}_i)^2} \tag{3.34}$$

Alle vorgestellten Distanzmaße wurden für die Klassifikation durch die Verwendung eines Schwellwertes entwickelt. Dieser wird oft als konstant angegeben. Damit ist eine Adaption auf verschiedene Geräuschumgebungen nicht möglich. [SMSG+01] beschreibt zwei Möglichkeiten, diesen Schwellwert adaptiv und damit störgeräuschunabhängig festzulegen. Der erste Vorschlag beruht auf einem Training auf die vorherrschende Geräuschumgebung mit Hilfe einer reinen Geräuschaufnahme und der Adaption des durch das Training ermittelten initialen Schwellwertes bei jedem weiteren Spracherkennungslauf. In [SMSG+01] werden mehrere Adaptionsvorschriften vorgestellt und diskutiert. Diese werden hier nicht erläutert, da das Verfahren ein vorhergehendes Training mit einer reinen Geräuschaufnahme benötigt, was nicht als adaptiv, sondern nur als „semi-adaptiv" angesehen wird.

Der zweite Vorschlag basiert auf der Annahme, dass in den ersten 200 Millisekunden einer Sprachaufnahme eine reine Geräuschaufnahme vorliegt. Dies entspricht in etwa den ersten 20 Analyserahmen und somit den ersten 20 cepstralen Merkmalsvektoren. Diese können zur Ermittlung eines initialen Schwellwertes herangezogen werden. Dadurch entfällt das Training aus dem ersten Vorschlag. Der Schwellwert kann dann während der Spracherkennung weiter adaptiert werden. Wird jedoch für die Dauer einer Äußerung ein konstantes stationäres und sich somit gering verändernde Störgeräusch angenommen, so kann auf die Adaption verzichtet werden, da der Schwellwert bei jeder Äußerung neu ermittelt wird.

Bereichskombinierte Sprach-Pause-Detektion

Der zweite Vorschlag in [SMSG+01] stellte den Ausgangspunkt für die Neuentwicklung eines gegenüber Luftfahrzeuggeräuschen robusten Sprach-Pause-Detektors in dieser Arbeit dar. Dabei wurde die Distanzberechnung zwischen aktuellem cepstralen Vektor und einem Störgeräusch-Mittelwertvektor auf bereits implementierte Verfahren der Spracherkennung zurückgeführt.

In [Vase96] wird erwähnt, dass sich geringfügig ändernde Störgeräusche, wie beispielsweise Helikopter-Geräusche, durch ein Hidden Markov Model (HMM) mit einem Zustand modelliert werden können. Das Training, die Auswertung der Gaußdichten sowie Hidden Markov Modelle werden in der Spracherkennungstechnologie verwendet und stehen somit in jedem Spracherkennungssystem zur Verfügung. Dies kann in der Signalvorverarbeitung ausgenutzt werden. Das Störgeräusch wird durch das Modell λ_{noise} mit einem Mittelwert μ_{noise} und einer diagonalen Kovarianzmatrix C_{noise} modelliert. Der Zustand des HMM wird durch ein akustisches Modell mit einem Codebuch mit einer Gaußdichte (kontinuierliche Ausgabeverteilung) und einer darüber definierten Verteilung (Mixturgewichte) beschrieben.

$$\begin{aligned} \lambda_{noise} &= \{\mu_{noise}, C_{noise}\} \\ N(x|\mu, C) &= \frac{1}{\sqrt{(2 \cdot \pi)^2 \cdot |C|}} \cdot e^{-\frac{1}{2} \cdot (x-\mu)^T \cdot C^{-1} \cdot (x-\mu)} \end{aligned} \tag{3.35}$$

Das HMM wird mit den ersten 20 Analyserahmen der Aufnahme trainiert. Danach kann jedem Analyserahmen durch Auswertung der Gaußdichte ein Wahrscheinlichkeitswert $p[n]$ zugewiesen werden. Die Klassifikation erfolgt dann durch den Vergleich

mit einem Schwellwert $v_{cepstThreshold}$. Dieser Schwellwert wird für jede zu verarbeitende Äußerung neu berechnet. Die Berechnungsvorschrift wurde in mehreren Versuchen im Rahmen dieser Arbeit ermittelt.

$$v_{cepstThreshold} = 0,8 \cdot \frac{1}{N} \cdot \sum_{i=0}^{N-1} p[i] \tag{3.36}$$

Vor dem Vergleich mit dem adaptiven Schwellwert wird die Kurve der für die gesamte Äußerung erhaltenen Wahrscheinlichkeitswerte mit einem Medianfilter geglättet, um kurzzeitige Signalspitzen (Peaks) zu beseitigen und damit Fehlklassifikationen zu vermeiden. Die Sprachsegmente des Endergebnisses werden um vier Rahmen vergrößert, um sicherzustellen, dass auch Verschlusslaute, wie beispielsweise „k", richtig detektiert werden und die Pausensegmente keine Sprachanteile enthalten.

Aus Gleichung 3.36 wird ersichtlich, dass diese Klassifikation nur funktioniert, wenn in der Äußerung Sprache vorhanden ist. Liegt eine reine Geräuschaufnahme vor, so kommt es aufgrund der Schwellwert-Berechnung zu Fehlklassifikationen. Dies könnte durch einen weiteren konstanten Schwellwert verhindert werden. Darauf wird verzichtet, da die Annahme zugrunde liegt, dass die Äußerungen durch Piloteninitiative (PTT[8]-Taste) aufgenommen werden und somit eine Verarbeitung von reinen Geräuschaufnahmen ausgeschlossen wird. Eine Fehlbedienung führt dann auch zu einer Fehlklassifikation und damit möglicherweise zu einer Fehlfunktion des Systems.

Die Annahme der Sprachfreiheit der ersten 20 Analyserahmen einer Äußerung muss nun etwas eingeschränkt werden. Sie wird zwar immer noch aufrecht erhalten. Allerdings können sich in den ersten 20 Rahmen Signalspitzen durch Störungen beim Einschalten der Aufnahme ansammeln, die zu einem Training falscher Störgeräuschdaten und damit zu einer Fehlklassifikation führen. Es sollte also adaptiv ein möglichst großes Trainingsintervall angegeben werden, auf dem das Störgeräusch-HMM trainiert werden kann. Dieses wird durch eine weitere Sprach-Pause-Detektion im Spektralbereich als Vorstufe zur Sprach-Pause-Detektion im cepstralen Bereich geleistet.

Dieser spektrale Sprach-Pause-Detektor ist einfach aufgebaut. Zunächst wird für jeden Analyserahmen die spektrale Energie berechnet. Diese so erhaltene „Energie-Kurve" wird mit Hilfe eines Medianfilters geglättet. Danach werden die Energiewerte auf das Intervall $[0 \ldots 1]$ normalisiert. Durch einen Vergleich mit einem Schwellwert $v_{specThreshold}$ wird die Klassifikation nach Sprach- oder Pauserahmen durchgeführt. Dieser Schwellwert wird adaptiv auf jede Äußerung angepasst. Hierzu wäre es von Vorteil, das Signal-zu-Rausch-Verhältnis (SNR) berechnen zu können. Dazu ist jedoch nach Gleichung 2.1 ein Sprach-Pause-Detektor erforderlich. Deshalb kann das SNR nicht für die Berechnung des Schwellwertes herangezogen werden. Statt dessen wird ein auf dem k-nächste-Nachbarn-Verfahren basierender, auf die gesamte Äußerung bezogener, geschätzter SNR-Wert (SNRk) verwendet. Diese Methode ist im Janus Recognition Toolkit (JRTk) enthalten und wird im Folgenden kurz vorgestellt.

Zur Ermittlung des SNRk-Wertes werden zunächst zwei Variablen $silMean$ und $speechMean$ mit der geringsten beziehungsweise höchsten durchschnittlichen Größenordnung (average Magnitude) des Signals initialisiert. Diese zwei Variablen werden als Startwerte für ein k-nächste-Nachbarn-Verfahren mit 10 Iterationen auf dem

[8]Push To Talk

Sprachsignal benutzt. Die so erhaltenen zwei Mittelwerte werden zur Berechnung des SNRk-Wertes verwendet.

$$SNRk = 20 \cdot (speechMean - silMean) \cdot \log_{10} 2,0 \tag{3.37}$$

Mit Hilfe des SNRk-Wertes konnte in dieser Arbeit in mehreren Versuchen eine adaptive Berechnungsvorschrift für den Schwellwert $v_{specThreshold}$ gefunden werden.

$$v_{specThreshold} = \frac{3,0}{SNRk} \tag{3.38}$$

Unter Verwendung dieses Schwellwertes wird eine Klassifikation durchgeführt. Durch die beschriebene Ermittlung des Schwellwerts wird auf den Entwicklungsdaten sichergestellt, dass die klassifizierten Pauserahmen keine Sprachanteile enthalten und somit das HMM richtig trainiert wird. Diese einfache Vorstufe ermittelt Intervalle zusammenhängender Pausenrahmen und stellt diese der cepstralen Sprach-Pause-Detektion zur Verfügung.

Durch einen Vergleich zwischen diesen Intervallen und der Anzahl der Merkmalsvektoren kann entschieden werden, ob eine Äußerung Sprache enthält oder nur Störgeräusche. Hierzu ist die Angabe eines konstanten Schwellwertes notwendig. Dadurch kann die Spracherkennung auch ohne Piloteninitiative erfolgen.

Es wird also eine, in dieser Arbeit entwickelte, zweistufige Sprach-Pause-Detektion mit adaptiven Schwellwerten im spektralen sowie dem cepstralen Bereich durchgeführt. Die Detektion im Spektralbereich ermittelt Pauserahmen. Diese werden als Intervalle mit Anfangs- und Endpunkt an die Sprach-Pause-Detektion im cepstralen Bereich weitergegeben. Diese trainiert auf diesen Bereichen ein HMM mit einem Zustand, mit dessen Hilfe danach die Wahrscheinlichkeit, dass ein Rahmen nur Störgeräusch enthält, für jeden Analyserahmen ermittelt wird. Aufgrund dieser Wahrscheinlichkeitswerte wird dann die Klassifikation aller Analyserahmen nach Sprach- oder Pauserahmen mit Hilfe eines weiteren adaptiven Schwellwertes durchgeführt. Aufgrund der Detektion im spektralen sowie im cepstralen Bereich wird das hier entwickelte Verfahren *bereichskombinierte Sprach-Pause-Detektion* genannt. In Abbildung 3.6 wird die Funktionsweise der bereichskombinierten Sprach-Pause-Detektion in der Standard-Vorverarbeitung nochmals verdeutlicht.

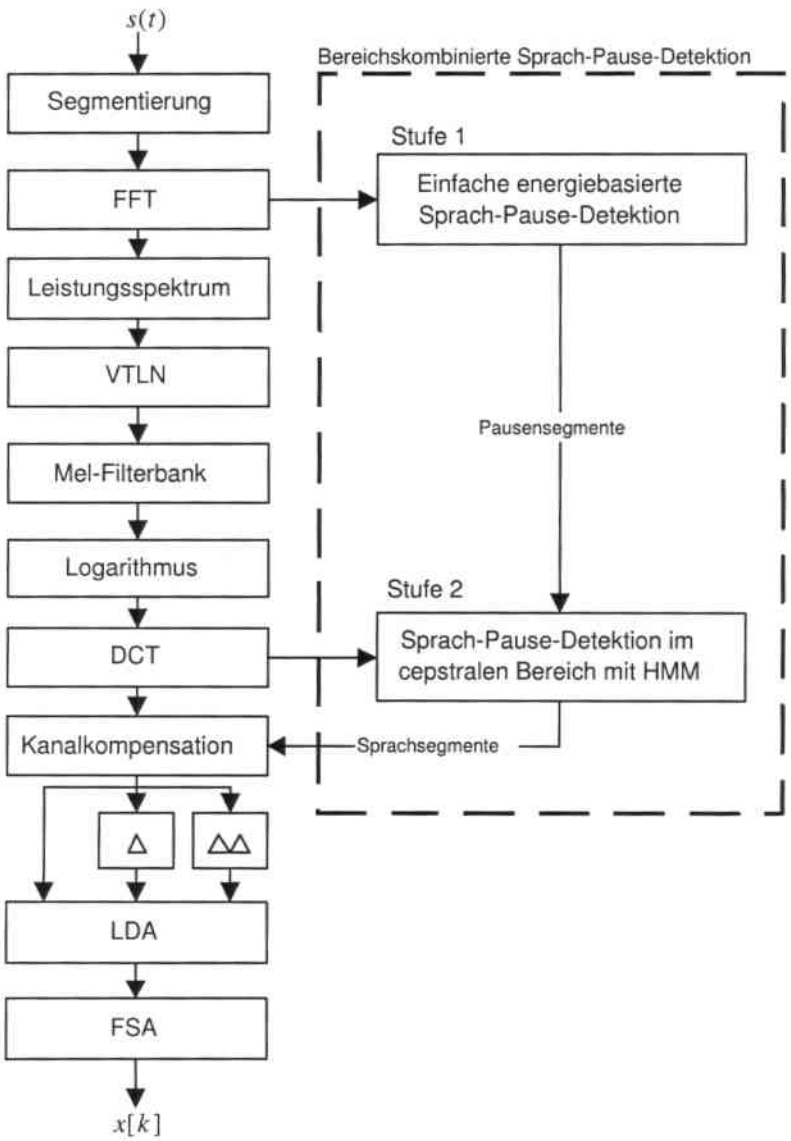


Abbildung 3.6: In die Standard-Vorverarbeitung integrierte bereichskombinierte Sprach-Pause-Detektion

4 Experimente und Ergebnisse

Eine Kompensation der Störgeräusche des Luftfahrzeugs sollen die in Kapitel 3 beschriebenen Verfahren leisten. Welche Kombination der vorgestellten Verfahren die beste Kompensation der Störgeräusche erreicht, wurde in zahlreichen Experimenten ermittelt. In diesen Experimenten wurde die bereichskombinierte Sprach-Pause-Detektion entwickelt und getestet. Zuerst werden die in dem Basis-Spracherkennungssystem verwendeten Akustiken und Sprachmodelle beschrieben. Danach folgt vor der Beschreibung der Experimente eine Übersicht über die zur Verfügung stehenden Sprachaufnahmen und die zur Beurteilung der Ergebnisse verwendeten Gütemaße.

4.1 Spracherkenner

Die verwendeten Spracherkennungssysteme wurden mit dem Janus Recognition Toolkit (JRTk) [FGHK+97] erstellt. Dieses wurde an den Interactive Systems Labs der Universität Karlsruhe (TH) sowie der Carnegie Mellon University Pittsburgh entwickelt. Das Toolkit enthält die gängigen Standardverfahren zur kontinuierlichen Spracherkennung mit Hidden Markov Modellen. In der aktuellen Version 5.0 wird der One-Pass-Decoder IBIS [SMFW01] verwendet, der an der Universität Karlsruhe (TH) entwickelt wurde. Dieser Decoder erlaubt neben den n-gram-basierten Sprachmodellen die Verwendung von kontextfreien Grammatiken. Verfahren zur Beschleunigung der Spracherkennung wie beispielsweise BBI[1] und Phonem-Lookahead wurden in dieser Arbeit nicht verwendet, da diese die Erkennungsraten negativ beeinflussen und zuerst geklärt werden sollte, ob überhaupt Spracherkennung in Luftfahrzeugen mit zufriedenstellenden Ergebnissen durchgeführt werden kann.

4.2 Trainingsdaten

Verschiedene Akustiken wurden für die Durchführung der Experimente verwendet. Die Akustiken basieren auf zwei bestehenden Corpora an Trainingsdaten: den Switchboard-Daten (SWB) und den Broadcast-News/English-Spontaneous-Scheduling-Task-Daten (BN/ESST). Wichtig ist, dass beide Corpora spontane

[1] Bucket Box Intersection Algorithm

sprachliche Äußerungen und ihre Besonderheiten, wie beispielsweise das Einfügen von „äh's", beinhalten. Während die Switchboard-Daten mit einer Abtastfrequenz von 8 kHz aufgenommen wurden, wurde bei den BN/ESST-Daten eine Abtastfrequenz von 16 kHz verwendet. Die Daten beider Corpora besitzen eine Quantisierung von 16 bit.

Die SWB-Daten bestehen aus spontanen Dialogen, aufgenommen über das Telefonnetz. Der Corpus beinhaltet circa 240 Stunden Sprache, gesprochen in vielen amerikanischen Dialekten von 500 Sprechern beider Geschlechter. Insgesamt wurden in den 240 Aufnahmestunden ungefähr 3 Millionen Wörter gesprochen. Der Corpus ist unterteilt in 2430 Dialoge mit einer durchschnittlichen Länge von 6 Minuten.

Die Corpora der BN/ESST-Daten werden nun getrennt beschrieben. Zunächst die Broadcast-News-Daten (BN). Sie bestehen aus Äußerungen von Nachrichtensprechern, auch bei Aussenaufnahmen, mit entsprechender Geräuschkulisse. Die Trainingsdaten umfassen ungefähr 106 Stunden Sprache, gesprochen durch 4019 Sprecher beider Geschlechter. Es wurden circa 1572436 Wörter gesprochen.

Die Daten des English-Spontaneous-Scheduling-Task (ESST) beinhalten spontane Sprache in Dialogen in der Domäne Termin- und Reiseplanung. Die Daten wurden mit Nahbesprechungsmikrofonen der Firma Sennheiser in einer geräuschfreien Laborumgebung aufgenommen. Die Sprachdaten von 242 Sprechern beider Geschlechter haben eine ungefähre Dauer von 35 Stunden.

4.3 Sprachmodelle

Vor dieser Arbeit wurde eine Umfrage unter Piloten bezüglich spontaner Anfragen an ein Flugnavigationssystem durchgeführt. Die Ergebnisse sind im Anhang A.1 aufgeführt. Auf der Grundlage der erhaltenen Antworten wurden Anfragen erstellt und im Sportflugzeug beziehungsweise Helikopter aufgenommen. Auf der Grundlage dieser Anfragen, die in Anhang A.5 aufgeführt sind, wurden die im Folgenden vorgestellten Sprachmodelle entwickelt. Für die Sprachaufnahmen aus dem Helikopter, die bereits 1999 aufgenommen wurden, wurden aufgrund von der Umfrage abweichender gesprochener Texte eigene Sprachmodelle aufgestellt. Alle Sprachmodelle wurden auf der Grundlage der Äußerungen der Entwicklungsdaten erstellt. Das verwendete Vokabular umfasst 405 Wörter.

Trigram-basiertes Sprachmodell (LM)

Das Trigram-Sprachmodell wurde aus den gesprochenen Texten der Entwicklungsdaten unter Verwendung des Interactice Systems LM Toolkit [RiSG97] erstellt. Das Sprachmodell gibt die Auftrittswahrscheinlichkeit einzelner Wörter sowie von Wortfolgen mit der maximalen Länge von drei Worten an und wird beim Spracherkennungsprozess als Wahrscheinlichkeit einer Wortfolge, gewichtet durch den Parameter lz, berücksichtigt. Die Perplexität der Trigram-Sprachmodelle beträgt auf den Texten der Testdaten für das Sportflugzeug 3,1, für den Helikopter 15,5.

Problematisch bei der Erstellung eines n-gram-Sprachmodells ist die benötigte Menge an Texten. Als zu wenig können die hier verwendeten Anfragen der Entwicklungsdaten, bestehend aus weniger als 1446 Wörtern, betrachtet werden. Durch dieses Vorgehen ist eine sehr gute Erkennungsrate bei den Entwicklungsdaten zu erwarten,

aber eine weit aus schlechtere Erkennungsrate bei unbekanntem Text, wie er teilweise in den Testdaten vorliegt. Aufgrund dieses Nachteils wurde das trigram-basierte Sprachmodell nur zum Vergleich erstellt. Es soll in einer weiteren Anwendung nicht mehr verwendet werden, da die Vorteile des auf einer kontextfreien Grammatik basierenden Sprachmodells überwiegen. Außerdem wurde bereits in anderen Domänen gezeigt, dass dieses Sprachmodell besser und der Erkennungsprozess schneller ist als bei der Verwendung eines n-gram-basierten Sprachmodells [FSSM+03].

Kontextfreie Grammatik (CFG)

Der im Spracherkennungssystem verwendete IBIS-Decoder ermöglicht die Verwendung einer kontextfreien Grammatik als Sprachmodell. Dadurch wird im Gegensatz zum n-gram-basierten Sprachmodell keine größere Menge an Text zur Erstellung des Sprachmodells benötigt. Die kontextfreie Grammatik kann von Hand erstellt werden. Dies wurde auf der Grundlage der Entwicklungsdaten durchgeführt. Allerdings kann die Grammatik durch ihren Aufbau wesentlich mehr Sätze erkennen, als in den Anfragen der Entwicklungsdaten angegeben sind. Hierzu ein Beispiel: Zur Erstellung des ngram-basierten Sprachmodells wurden nur die im Text angegebenen Funkfrequenzen benutzt, in der Grammatik wurden diese als

$$\begin{aligned} \text{[frequency]} &\longrightarrow \text{[number][number][number] decimal [number][number]} \\ \text{[number]} &\longrightarrow 0|\ldots|9 \end{aligned}$$

modelliert, so dass die Grammatik die Struktur der Frequenzen des Flugfunks berücksichtigt. Deshalb wird eine etwas schlechtere Erkennungsrate auf den Entwicklungsdaten, dafür aber eine bessere Erkennungsrate auf den Testdaten erwartet. Weiterhin kann die Ausgabe des Spracherkenners durch die Möglichkeit der Weiterverarbeitung durch einen Parser zum Beispiel in einem Dialogsystem problemlos verwendet werden.

4.4 Test-/Entwicklungsdaten

Sprachdaten aus einem Helikopter sowie einem Sportflugzeug standen zur Verfügung. Diese wurden im Rahmen dieser Arbeit aufgenommen oder standen bereits dem Institut für Logik, Komplexität und Deduktionssysteme der Universität Karlsruhe (TH) zur Verfügung. Einen Überblick über die verwendeten Daten gibt Tabelle 4.1. Insgesamt standen 1629 Äußerungen zur Verfügung. Davon wurden 214 Äußerungen von drei männlichen Sprechern im Helikopter und 1199 Äußerungen von einem männlichen Sprecher im Sportflugzeug aufgenommen. Diese Daten wurden in Entwicklungs- und Testdaten aufgeteilt. Als Testdaten standen insgesamt 659 Äußerungen zur Verfügung. Davon wurden 144 Äußerungen im Helikopter und 299 Äußerungen im Sportflugzeug aufgenommen. Die restlichen 970 Luftfahrzeugdaten standen für die Entwicklung zur Verfügung. Eine genaue Beschreibung der Durchführung der Aufnahmen ist im Anhang A.2 gegeben. Die im Anhang erwähnten Vergleichsaufnahmen mit den gleichen Äußerungen vom gleichen Sprecher aus dem Sportflugzeug wurden den Helikopter-Testdaten zugeschlagen. In den insgesamt 1199 Sprachaufnahmen aus dem Sportflugzeug wurden nicht 1199 verschiedene Texte gesprochen, sondern die in Anhang A.5 angegebenen 214 Anfragen bestehend

aus 1446 Wörtern wurden in verschiedenen Flugzuständen, zum Beispiel Rollen am Boden, Start, Steig-, Sink- und Reiseflug, wiederholt gesprochen. Dies machte die Aufnahme sehr vieler unterschiedlicher Störgeräusche möglich und deckte den ganzen Einsatzbereich des Sportflugzeugs ab. Bei den Aufnahmen im Helikopter wurde auf dieses Vorgehen verzichtet, da eine Unterteilung in Flugphasen nur Anlassen und Fliegen unterscheiden würde. Es wurden somit weniger Sprachdaten aufgenommen. Die Äußerungen wurden deshalb nicht wiederholt, sondern unterscheiden sich. Allerdings ensprechen die mit dem Sennheiser-Headset gemachten Aufnahmen den Anfragen im Anhang. Die mit dem Bose-Headset durchgeführten Aufzeichnungen enthalten Städtenamen, ICAO[2]-Codes und Anfragen an ein Navigationssystem. Da das Spracherkennungssystem durchaus auch im Luftfahrzeug ohne laufende Motoren eingesetzt werden soll, wurden für die Testdaten im Labor alle im Anhang A.5 aufgeführten Anfragen aufgenommen. Alle Daten haben eine Gesamtdauer von ungefähr 76 Minuten, davon entfallen circa 15 Minuten auf die Helikopter-Sprachdaten und circa 49 Minuten auf die Daten aus dem Sportflugzeug.

Set	Luftfahrzeug	Headset	Äußerungen	Sprecher	Minuten
Entwicklungs-daten	Helikopter	Bose	70	2	3,5
	Sportflugzeug	Sennheiser	900	1	37,0
Testdaten	Helikopter	Bose	30	2	1,5
		Sennheiser	114	1	10
	Sportflugzeug	Sennheiser	299	1	12,0
	Labor	Sennheiser	216	1	12,0

Tabelle 4.1: Aufteilung der Sprachdaten

4.5 Verwendete Gütemaße

Zur Beurteilung der Erkennungsleistung des Spracherkennungssystems sowie von Verbesserungen der Erkennungsleistung wurden als Gütemaße Erkennungsraten sowie die relative Fehlerreduktion verwendet.

Ein häufig verwendetes Maß ist die Wortfehlerrate (WFR). Sie gibt das Verhältnis der Anzahl von falsch erkannten Wörtern N_{err} zu der Gesamtanzahl der Wörter N an. Die falsch erkannten Wörter setzen sich aus Auslasse- und Einfügefehlern (deletions, insertions) sowie verwechselten Worten (substitutions) zusammen.

$$WFR = \frac{N_{err}}{N} = \frac{N_{del} + N_{ins} + N_{sub}}{N} \tag{4.1}$$

Die Worterkennungsrate oder Wortakkuratheit (WA) ist der verbleibende Teil und ist definiert als

$$WA = 100\% - WFR = \frac{N - N_{err}}{N} = \frac{N - N_{del} - N_{ins} - N_{sub}}{N} \tag{4.2}$$

Weiterhin wird noch die Satzfehlerrate (SFR) angegeben. Sie wird berechnet als

[2]International Civil Aviation Organization

$$SFR = 100\% - SA \tag{4.3}$$

wobei die Satzakkuratheit (SA) das Verhältnis der korrekt erkannten Äußerungen zu der Gesamtanzahl der Äußerungen angibt.

Werden bestehende Systeme oder Verfahren geändert, so sind absolute Fehlerraten nicht sehr aussagekräftig, da beispielsweise eine Verbesserung der Wortfehlerrate von 5% auf 4% signifikanter ist als eine Verbesserung von 45% auf 44%. Deshalb wird die relative Fehlerreduktion (RFR) als Verhältnis der Differenz der Fehlerraten von altem System (A) und modifiziertem System (B) zu der Fehlerrate des alten Systems (A) definiert.

$$RFR = \frac{WFR_A - WFR_B}{WFR_A} \cdot 100\% \tag{4.4}$$

Der Real Time Factor (RTF) gibt an, um welchen Faktor sich die Bearbeitungszeit, die Zeit bis die Ausgabe des Spracherkenners bereitgestellt wird, von der Länge der Sprachaufnahme unterscheidet. Bei einem RTF von eins entspricht die Bearbeitungszeit genau der Länge der Sprachaufnahme. Ist der RTF zwei, so benötigt das Spracherkennungssystem doppelt so lange als die Äußerung (Eingabe in das System) dauert. Dieser wurde für das erstellte Spracherkennungssystem mit den besten Vorverarbeitungsverfahren ermittelt.

4.6 Experimente und Ergebnisse

In diesem Abschnitt werden die Experimente mit den in Kapitel 3 erläuterten Verfahren beschrieben. Zuerst wurde der Sprach-Pause-Detektor der Standard-Vorverarbeitung an die geänderte Geräuschumgebung in den Luftfahrzeugen angepasst. Der Sprach-Pause-Detektor ist für die Kanalkompensation durch die sprachbasierte CMS, die zur Standard-Vorverarbeitung gehört, erforderlich. Es wurde ein Spracherkenner mit der BN/ESST-Akustik verwendet. Diese besitzt ungefähr 2143 Codebücher mit 16 Codebuchvektoren mit einer Dimension von 32. Mit diesem Spracherkenner wurden mehrere Erkennungsläufe auf den Entwicklungsdaten des Helikopters sowie des Sportflugzeugs durchgeführt. In diesen Erkennungsläufen wurden die Parameter lz, lp, das filler-penalty in Abhängigkeit zum verwendeten Sprachmodell festgelegt. Durch diese Festlegung werden die weiteren Verfahren eingeschränkt, da ihre maximale Erkennungsleistung bei unterschiedlichen Einstellungen der Parameter lz und lp erreicht wird. Die Schwankungen der Erkennungsraten waren jedoch kleiner als fünf Prozent, so dass für die Auswahl eines für die Spracherkennung im Luftfahrzeugcockpit geeigneten Verfahrens dies in Kauf genommen und kein lattice-rescoring verwendet wurde. Die festgelegten Werte sind in Tabelle 4.2 dargestellt. Die Erkennungsergebnisse mit diesen festgelegten Parametern sind in Tabelle 4.3 dargestellt und dienen als Basis für die Angabe der relativen Fehlerreduktion bei allen weiteren Versuchen. Diese Parametereinschränkung wurde nur für die Experimente auf den Entwicklungsdaten gemacht. Bei den abschließenden Versuchen auf den Testdaten wurde ein lattice-rescoring verwendet und dadurch die besten Werte für die Parameter lz und lp ermittelt.

Wie erwartet sind die Ergebnisse des trigram-basierten Sprachmodells besser als die des Sprachmodells mit kontextfreier Grammatik innerhalb einer Luftfahrzeugklasse

Sprachmodell	lz	lp	filler-penalty
CFG	20	4	30
LM	38	4	30

Tabelle 4.2: Festgelegte Parameter des Spracherkenners

Verfahren	Luftfahrzeug	CFG	LM
Standard-Vorverarbeitung	Helikopter	WFR 33,74% SFR 61,43%	WFR 31,78% SFR 60,00%
	Sportflugzeug	WFR 22,61% SFR 48,33%	WFR 10,83% SFR 26,67%

Tabelle 4.3: Erkennungsergebnisse mit der Standard-Vorverarbeitung

– Helikopter oder Sportflugzeug. Für die erheblichen Unterschiede zwischen den Luftfahrzeugklassen können zwei Gründe angegeben werden:

1. Die Störgeräusche sind beim Helikopter durch die Turbine und den Rotor in einem sehr größeren spektralen Bereich vorzufinden als die niederfrequenten Störungen des Kolbenmotors im Sportflugzeug. Auch die Verwendung unterschiedlicher Headsets wirkt sich auf die Erkennungsergebnisse aus. Nähere Angaben wurden hierzu bereits im Kapitel 2.1 gemacht.

2. Bei einer näheren Betrachtung der Ausgaben des Spracherkenners wurde festgestellt, dass in der Grammatik öfter ein falscher Abstieg im aufgebauten Baum gewählt wird. Durch die Wahl des falschen Weges ist dadurch der ganze Satz falsch, was die erheblichen Unterschiede in der Wortfehlerrate zwischen Grammatik und trigram-basiertem Sprachmodell begründet. Auch der Aufbau der Grammatik und die damit zur Verfügung stehenden Auswahlmöglichkeiten pro Wort führen zu Substitutionsfehlern.

Bereichskombinierte Sprach-Pause-Detektion

Westphal verwendete in [West01] für die MAM einen einfachen energiebasierten Sprach-Pause-Detektor, wie er in ähnlicher Weise in der Standard-Vorverarbeitung benutzt wird. Dieser wurde mit Hilfe eines Schwellwertes auf die Störgeräusche der Autoumgebung eingestellt und zunächst auf die Luftfahrzeugdaten angewandt. Es stellte sich jedoch heraus, dass für die Daten aus dem Helikopter und dem Sportflugzeug unterschiedliche Schwellwerte verwendet werden mussten. Dies bedeutete eine Abhängigkeit des Detektors vom verwendeten Luftfahrzeug und dessen Geräuschumgebung. Da diese Abhängigkeit vermieden werden sollte, wurde die bereichskombinierte Sprach-Pause-Detektion (siehe Kapitel 3.6) entwickelt.

Zuerst wurde eine Entscheidung für den cepstralen Bereich getroffen, da dieser gegenüber dem spektralen Bereich eine hohe Unabhängigkeit von Amplitude und Störgeräuschlevel bietet. Es wurden aufgrund der Detektionskurven, die in den Abbildungen 4.1 bis 4.4 dargestellt sind, zunächst eine subjektive Entscheidung getroffen. Der integrale Ansatz liefert gute Ergebnisse, enthält jedoch ein hohes Grundrauschen, das

sich eventuell negativ auf die Schwellwertberechnung ausüben kann. Der euklidische Ansatz zeigt gleich gute Ergebnisse, enthält jedoch ein geringeres Grundrauschen. Die Wahrscheinlichkeiten der Auswertung der Gaußdichte des HMM zeichnen sich durch einen hohen „Gleichstromanteil" aus, der allerdings durch eine Subtraktion entfernt werden kann. Aufgrund dieser Beobachtungen wurde eine Entscheidung zugunsten des HMM-Ansatzes und des Ansatzes mit der euklidischen Distanz getroffen.

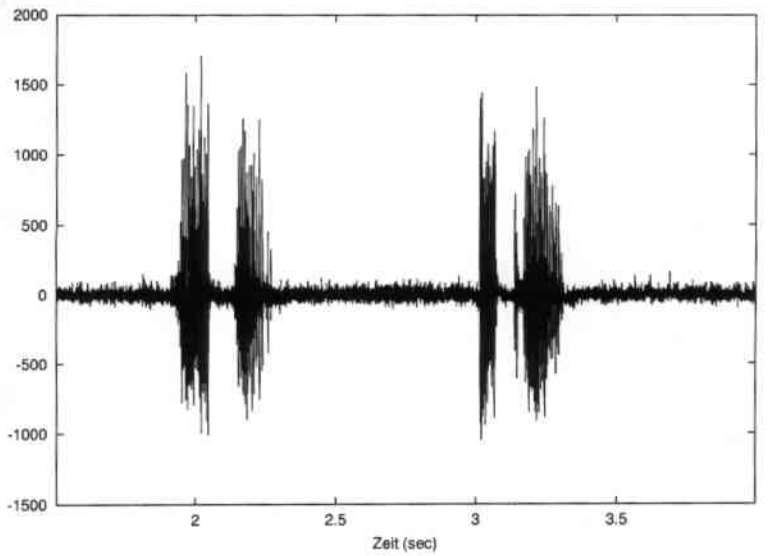


Abbildung 4.1: Sprachaufnahme „papa papa", Gesamtdauer 4,19sec

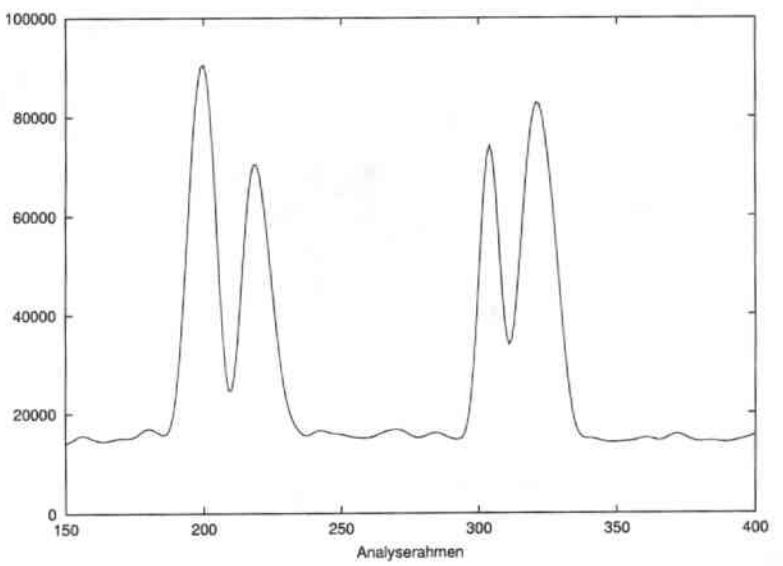


Abbildung 4.2: Wahrscheinlichkeitswerte der Auswertung der Gaußdichte des HMM

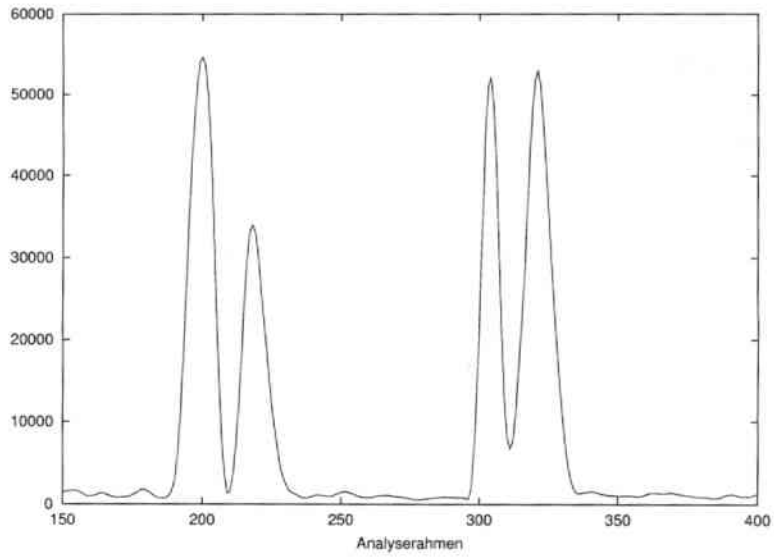


Abbildung 4.3: Euklidisches Distanzmaß

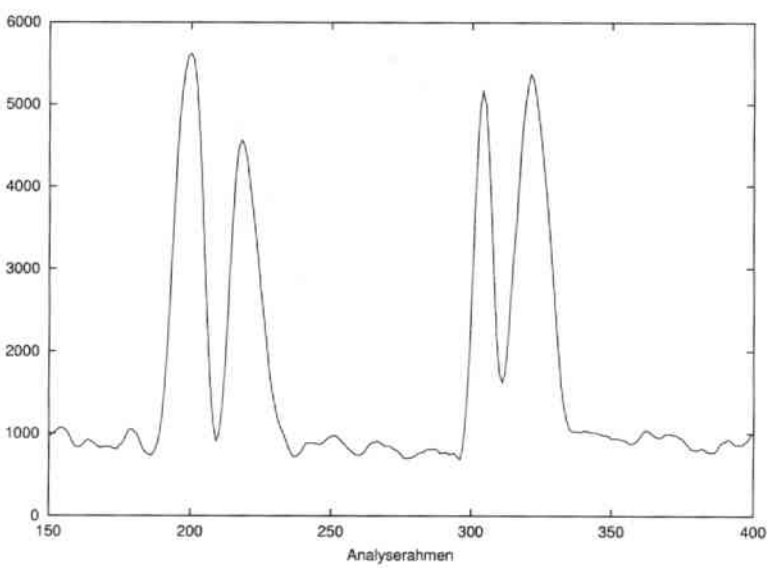


Abbildung 4.4: Integrales Distanzmaß

Getestet wurden ein cepstraler Sprach-Pause-Detektor mit euklidischem Distanzmaß und mit einem HMM. Hierzu wurde das euklidische Distanzmaß abweichend von Gleichung 3.32 implementiert. Da das JRTk ein Framework zur Vektorberechnung zur Verfügung stellt, wurde die Distanz des aktuellen Merkmalsvektors $\vec{x}[n]$ zum vorher ermittelten Mittelwertvektor $\bar{\vec{x}}$ wie in Gleichung 4.5 angegeben berechnet. Auf die Normierung des Distanzwertes wurde verzichtet, da diese aus nur einer Multiplikation mit einem konstanten Faktor besteht.

$$d[n] = \left\| \vec{x}[n] - \bar{\vec{x}} \right\|^2 \tag{4.5}$$

Als Testumgebung wurde die Segmentierung von Telefon-Sprachdaten verwendet. Hierfür standen verschiedene, zum Teil inoffizielle Erkennungsergebnisse mit unterschiedlichen Segmentierungen zur Verfügung, so dass die Sprach-Pause-Detektion entsprechend des resultierenden Erkennungsergebnisses eingestuft werden konnte. Bei diesen Versuchen zeigte sich jedoch schnell, dass in den zuerst angenommenen ersten 20 Geräuschanalyserahmen die schon erwähnten – durch den Aufnahmebeginn verursachten – Signalspitzen störend auf die Sprach-Pause-Detektion wirkten und das Erkennungsergebnis verschlechterten. Deshalb wurde eine Lösung gesucht, das Training der Modelle bei beiden Verfahren zu verbessern. Die in der Standard-Vorverarbeitung verwendete Sprach-Pause-Detektion im spektralen Bereich bot sich dafür an, es musste nur eine adaptive Berechnungsvorschrift für den Schwellwert gefunden werden. Geeignet erschien eine Abhängigkeit vom SNR der Äußerung. Dieses wird näherungsweise durch die Funktion SNRk – beschrieben in Kapitel 3.6 – erfüllt. Damit konnte nach weiteren Experimenten eine entsprechende Berechnungsvorschrift aufgestellt und implementiert werden.

Die zur Verfügung stehenden Vergleichsergebnisse wurden durch mehrere Schritte erzeugt. Es wurden zuerst mit Hilfe eines energiebasierten Sprach-Pause-Detektors mit konstanten Schwellwerten die Analyserahmen in drei Kategorien unterteilt: Sprache, Pause und Unsicher. Auf den Sprach- beziehungsweise Pausensegmenten werden zwei GMM-Modelle trainiert. Diese werden verwendet, um die „Unsicher"-Segmente zu klassifizieren. Dieses Training der Modelle mit anschließender Klassifikation wird mehrmals wiederholt. Nach jedem Klassifikationsschritt werden kurze, falsch detektierte Sprachsegmente durch einen Vergleich mit einer weiteren Konstanten zu Pausensegmenten umgewidmet. Dies wurde in den Segmentierungen mit HMM und euklidischer Distanz auch mit Hilfe von Konstanten durchgeführt. Die Vergleichssegmentierungen erreichten bei mehreren Erstellungs-Iterationen eine Wortfehlerrate von 46,3% und 46,4%. Die Segmentierung mit den Hidden-Markov-Modellen erreichte eine Wortfehlerrate von 47,7%, die mit euklidischer Distanz 49,5%. Beide Segmentierung wurden ohne Iteration erstellt. Aus diesen Ergebnissen ist zu sehen, dass die Sprach-Pause-Detektion mit euklidischer Distanz schlechter abschneidet als die mit HMM. Im Vergleich mit den anderen Verfahren liegen beide jedoch hinter den anderen Verfahren, was auf die iterative Durchführung dieser zurückgeführt wird. Die Segmentierung mit Sprach-Pause-Detektion mit Hidden Markov Model wird aufgrund dieser Ergebnisse als gut angesehen. Es wurden beide Ansätze weiter verfolgt und in die Standard-Vorverarbeitung integriert und getestet. Das Ergebnis ist in Tabelle 4.4 dargestellt.

Aufgrund der dargestellten Ergebnisse wurde eine Entscheidung zugunsten der Sprach-Pause-Detektion mit HMM getroffen. Diese wurde in Kapitel 3.6 als be-

Verfahren	Luftfahrzeug	Maß	CFG	LM
Standard-Vorverarbeitung mit cepstraler Sprach-Pause-Detektion	Helikopter	HMM	WFR 29,34% SFR 51,43%	WFR 26,16% SFR 52,86%
		Euklid	WFR 28,61% SFR 57,14%	WFR 29,83% SFR 61,43%
	Sportflugzeug	HMM	WFR 20,98% SFR 45,22%	WFR 8,74% SFR 21,67%
		Euklid	WFR 21,45% SFR 44,44%	WFR 8,98% SFR 22,44%

Tabelle 4.4: Erkennungsergebnisse mit der Standard-Vorverarbeitung mit cepstraler Sprach-Pause-Detektion

reichskombinierte Sprach-Pause-Detektion vorgestellt. Die Entscheidung verstärkte ein Test zur Schnelligkeit der Detektion. Dabei wurde die Sprach-Pause-Detektion auf 83 Sprachaufnahmen mit einer Gesamtdauer von 400,67 Sekunden angewandt. Die Detektion mit dem HMM benötigte dazu 44,98 Sekunden, die Detektion mit euklidischer Distanz 72,88 Sekunden. Die Entscheidung für das HMM stellte sich damit als richtig heraus. Damit hat die Detektion mit HMM einen rechnerischen RTF von 0,11. Dabei ist zu beachten, dass die Implementierung der Detektion mit HMM im Hinblick auf die Geschwindigkeit noch optimiert werden kann. Die Implementierung der Detektion mit euklidischer Distanz ist dagegen schon optimal.

Die bereichskombinierte Sprach-Pause-Detektion wurde in allen weiteren Verfahren angewendet. Als nächstes Verfahren wurde die von Westphal entwickelte modellkombinationsbasierte akustische Transformation betrachtet.

Modellkombinationsbasierte Akustische Transformation

Für den Einsatz der modellkombinationsbasierten akustischen Transformation (MAM) war ein Neutraining des Modells ungestörter Sprache auf dem BN/ESST-Corpus erforderlich. Dieses Modell besitzt 100 Gaußdichten. In der von Westphal implementierten Form ergab die modellkombinationsbasierte akustische Transformation die in Tabelle 4.5 dargestellten Erkennungsergebnisse, die eine Verbesserung gegenüber der Standard-Vorverarbeitung zeigen. Weiterhin wurde die bereichskombinierte Sprach-Pause-Detektion auch in der MAM angewendet, wozu wenige Veränderungen des Verfahrens notwendig waren. Durch die Einführung der sprachbasierten cepstralen Mittelwertsubtraktion mit einer weiteren, von der MAM getrennten bereichskombinierten Sprach-Pause-Detektion wurde eine weitere Steigerung der Erkennungsraten (Tabelle 4.5) erzielt. Diese Modifikationen sind in Abbildung 4.5 dargestellt, die erreichten relativen Fehlerreduktionen in Tabelle 4.6.

Spektrale Subtraktion mit Geräuschschätzung durch Minimum Statistik

In mehreren Versuchen auf den Entwicklungsdaten wurden die Parameter der Störgeräuschschätzung durch Minimum Statistik und der spektralen Subtraktion auf den Entwicklungsdaten aus dem Helikopter bestimmt. Die spektrale Subtraktion wurde in die Standard-Vorverarbeitung eingebettet. Zur Kanalkompensation wurde die

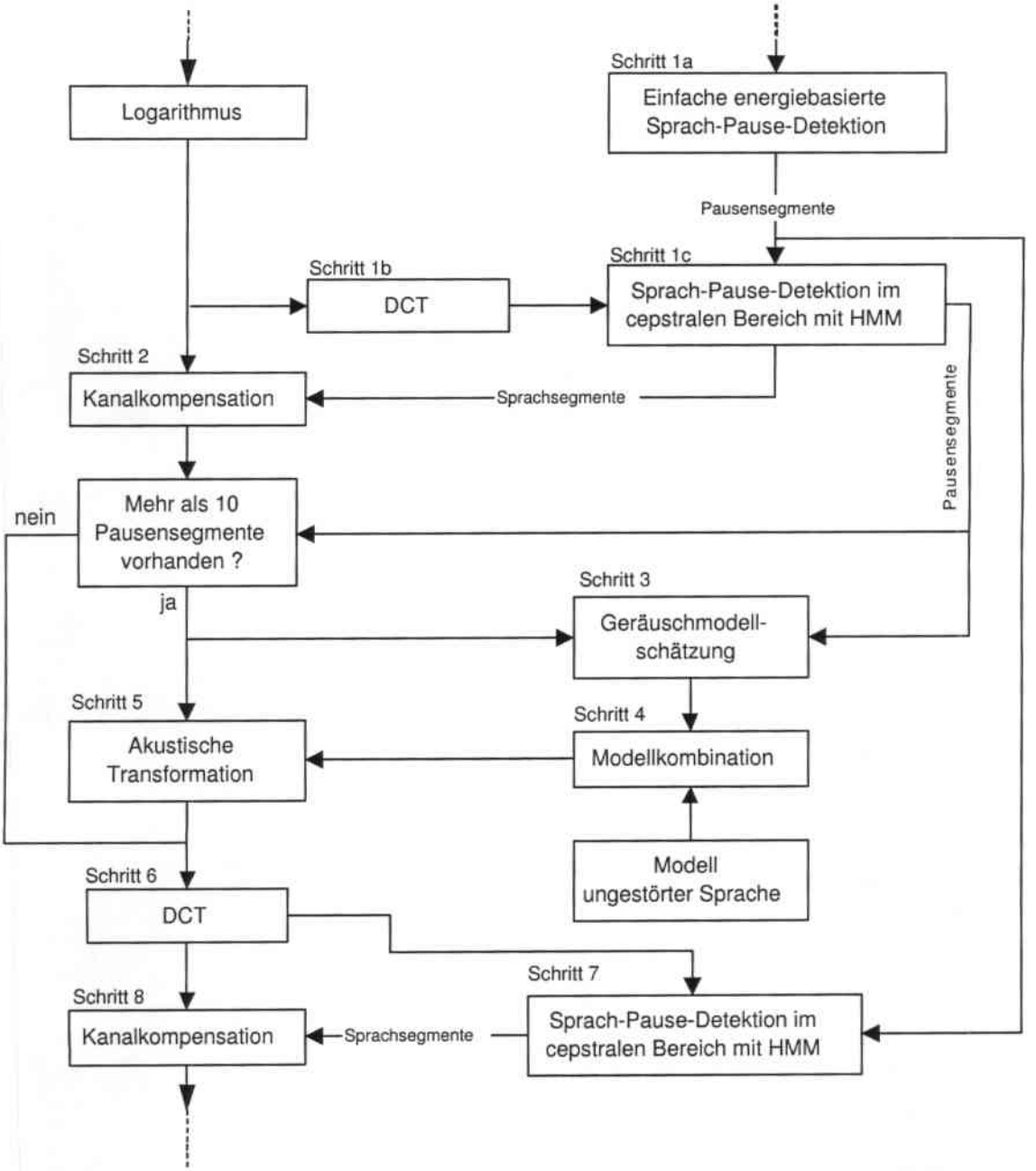


Abbildung 4.5: modifizierte MAM

Verfahren	Luftfahrzeug	CFG	LM
MAM	Helikopter	WFR 28,61% SFR 51,43%	WFR 21,27% SFR 47,14%
	Sportflugzeug	WFR 19,52% SFR 41,33%	WFR 8,48% SFR 21,67%
MAM mit bereichskombinierter Sprach-Pause-Detektion	Helikopter	WFR 26,89% SFR 47,14%	WFR 19,80% SFR 44,29%
	Sportflugzeug	WFR 17,86% SFR 39,22%	WFR 7,55% SFR 19,11%
MAM mit SCMS und zweifacher bereichskombinierter Sprach-Pause-Detektion	Helikopter	WFR 24,69% SFR 48,57%	WFR 19,07% SFR 45,71%
	Sportflugzeug	WFR 16,77% SFR 36,33%	WFR 6,99% SFR 17,89%

Tabelle 4.5: Erkennungsergebnisse mit der modellkombinationsbasierten akustischen Transformation

Modifikation	Luftfahrzeug	CFG	LM
MAM $\longrightarrow$ MAM mit SCMS und zweifacher bereichskombinierter Sprach-Pause-Detektion	Helikopter	RFR 13,70%	RFR 10,34%
	Sportflugzeug	RFR 14,09%	RFR 17,57%

Tabelle 4.6: Erreichte relative Fehlerreduktionen der MAM-Modifikationen

sprachbasierte CMS mit der bereichskombinierten Sprach-Pause-Detektion verwendet. Die Ergebnisse der Erkennungsläufe mit den vorher festgelegten Parametern sind in Tabelle 4.7 widergegeben.

Verfahren	Luftfahrzeug	CFG	LM
Spektrale Subtraktion mit Störgeräuschschätzung durch Minimum Statistik	Helikopter	WFR 21,76% SFR 47,14%	WFR 22,25% SFR 50,00%
	Sportflugzeug	WFR 20,02% SFR 43,67%	WFR 8,19% SFR 20,78%

Tabelle 4.7: Erkennungsergebnisse mit der Störgeräuschschätzung und spektralen Subtraktion

Die Ergebnisse der spektralen Subtraktion sind größtenteils schlechter als die Ergebnisse der MAM. Nur im Helikopter mit der Verwendung der kontextfreien Grammatik ist eine Verbesserung vorhanden. Dies ist durch die Parameterbestimmung bedingt. Die schlechten Ergebnisse konnten auf die Störgeräuschschätzung zurückgeführt werden. Diese verlangt für jede Geräuschänderung unterschiedliche Parameter. Da die Aufnahmen des Sportflugzeugs alle Flugphasen umfassen, variiert dort das Störgeräusch viel mehr als bei den Aufnahmen im Helikopter, welche bei laufender Turbine und Rotor aufgenommen wurden. Trotz sorgfältiger Bestimmung der konstanten Parameter der Störgeräuschschätzung zeigt sich dieses Vorgehen als nicht praktikabel.

Das durch die Störgeräuschschätzung ermittelte Störgeräusch wird direkt vom Spektrum des Signals der Äußerung abgezogen. Diese Schätzung ist, wie oben beschrieben, sehr fehleranfällig. Das Verfahren kann verbessert werden, wenn das Störgeräusch nicht geschätzt werden muss, sondern es bereits in einem zweiten Kanal mit der Äußerung aufgenommen wird. Dies wurde bei den Aufnahmen aus dem Sportflugzeug gemacht. Das Störsignal wurde bei der Verarbeitung in den Spektralbereich transformiert, danach wurde es nach den Gleichungen 3.7 und 3.8 zeitlich und in der Energie der einzelnen Analyserahmen angepasst. Die Anpassungsschritte wurden implementiert und mit Hilfe zweier gleicher Signale (Autokorrelation) auf ihre Korrektheit überprüft. Das erhaltene angepasste Störgeräuschspektrum wurde vom Spektrum der Äußerung subtrahiert. Weiterhin wurde das angepasste Störgeräuschspektrum in der MAM verwendet. Beide Verfahren zeigten jedoch schlechtere Ergebnisse als die Standard-Vorverarbeitung. Dies war sehr überraschend, da eine Verbesserung der Worterkennungsraten erwartet wurde. Deshalb wurde die Aufnahmesituation mit den entsprechenden Geräten in einer Laborumgebung nachgestellt und überprüft. Dabei zeigte sich aufgrund der sehr starken Richtwirkung der verwendeten Headset-Mikrofone, dass die erhaltenen Signale sich ab einem räumlichen Abstand von circa 10 cm der Mikrofone sehr unterscheiden und nicht korrelierbar sind. Dieser Effekt wird im Luftfahrzeug durch die unterschiedliche Akustik innerhalb der Luftfahrzeugzelle nochmals verstärkt. Somit konnten die zweikanaligen Aufnahmen in den Versuchen nicht genutzt werden.

Neutraining der BN/ESST-Akustik

Alle bisherigen Versuche wurden mit der bereits vorhandenen und mit einem einfachen spektralen Sprach-Pause-Detektor trainierten BN/ESST-Akustik durchgeführt. Die aus den Vorverarbeitungen resultierenden Merkmalsvektoren unterscheiden sich dabei von den verwendeten Trainingsvektoren um die Größe der Bereiche, auf denen die sprachbasierte CMS durchgeführt wird, da in den Testläufen die bereichskombinierte Sprach-Pause-Detektion anstatt des einfachen spektralen Sprach-Pause-Detektors verwendet wurde. Um diese Annahmen zu verifizieren wurde die Akustik mit der bereichskombinierten Sprach-Pause-Detektion auf dem BN/ESST-Corpus neu trainiert. Dieses Neutraining lässt allerdings nicht allzu große Steigerungen der Erkennungsraten erwarten, da die während des Trainings subtrahierten Mittelwerte vor dem Training für jeden Sprecher bestimmt werden. Auch hierzu wurde die bereichskombinierte Sprach-Pause-Detektion eingesetzt. Nach dem Neutraining wurden die spektrale Subtraktion sowie die MAM mit SCMS und bereichskombinierter Sprach-Pause-Detektion auf den Entwicklungsdaten getestet. Wie der Tabelle 4.8 zu entnehmen ist, resultierte das Neutraining bei Verwendung der MAM in der Luftfahrzeugklasse „Sportflugzeug" in einer geringen Verringerung sowohl der Wortfehlerrate als auch der Satzfehlerrate. In der Luftfahrzeugklasse „Helikopter" konnte überraschend keine Steigerung der Wortakkuratheit festgestellt werden. Bei der spektralen Subtraktion nahmen die Wortfehlerraten zum Teil drastisch zu. Dies hat folgenden Grund: Die zum Training verwendeten Daten bestehen aus zum Teil sehr langen Segmenten. Es konnte durch Versuche festgestellt werden, dass die Korrektheit der bereichskombinierten Sprach-Pause-Detektion aufgrund der Schwellwertberechnung im cepstralen Bereich bei langen Äußerungen abnimmt. Da die Sprachaufnahmen aus dem Sportflugzeug im Durchschnitt länger sind als die Aufnahmen aus dem Helikopter, könnte die geringe Verbesserung der Sportflugzeug-Worterkennungsraten damit begründet werden.

Die bereichskombinierte Sprach-Pause-Detektion ist aufgrund der gemachten Ausführungen im Training nur bedingt zu gebrauchen. Ihr Einsatz sollte auf bereits trainierte Spracherkennungssysteme beschränkt bleiben. Die drastische Erhöhung der Wortfehlerrate bei der spektralen Subtraktion kann wiederum mit den suboptimal festgelegten Parametern der Störgeräuschschätzung begründet werden.

Verfahren	Luftfahrzeug	CFG	LM
MAM mit SCMS und zweifacher bereichskombinierter Sprach-Pause-Detektion	Helikopter	WFR 30,56% SFR 54,29%	WFR 21,76% SFR 50,00%
	Sportflugzeug	WFR 16,42% SFR 36,67%	WFR 6,79% SFR 17,22%
Spektrale Subtraktion mit Störgeräuschschätzung durch Minimum Statistik	Helikopter	WFR 41,56% SFR 70,00%	WFR 23,23% SFR 44,29%
	Sportflugzeug	WFR 20,93% SFR 45,22%	WFR 8,26% SFR 20,67%

Tabelle 4.8: Erkennungsergebnisse nach dem Neutraining der BN/ESST-Akustik mit der bereichskombinierten Sprach-Pause-Detektion

Kanalkompensation

Die Kanalkompensation wurde in allen Versuchen durch die sprachbasierte cepstrale Mittelwertsubtraktion, die auch in der Standard-Vorverarbeitung verwendet wird, durchgeführt. In Kapitel 3.5 wurden noch weitere auf der sprachbasierten Mittelwertsubtraktion basierende Verfahren vorgestellt. Davon wurde die 2CMS aufgrund der möglichen Überlagerung von Sprach- und Pausevektoren bereits ausgeschlossen. Deshalb wurden damit keine Versuche durchgeführt. Da aufgrund der Verwendung von Nahbesprechungsmikrofonen die sprachbasierte cepstrale Mittelwertsubtraktion zur Kanalkompensation als ausreichend angesehen wird, wurde auch die 2DCMS nicht implementiert und in Versuchen betrachtet. Weiterhin stellte Westphal in [West01] fest, dass die Fehlerratenreduktion der 2DCMS durch die MAM übertroffen wird.

Switchboard-Akustik

Bis jetzt wurde der Übertragungsbereich der verwendeten Headset-Mikrofone nicht beachtet. Da der Übertragungsbereich fast dem Bereich eines analogen Standard-Telefonkanals entspricht, liegt es nahe, die Verfahren mit der Switchboard-Akustik zu testen. Dieses System wurde mit Telefon-Sprachdaten mit einer Abtastrate von 8 kHz trainiert, so dass die Sprachaufnahmen in der Vorverarbeitung zusätzlich in der Abtastrate reduziert und an die Bandbreite des Telefonkanals angepasst werden mussten. Diese Akustik besitzt 2000 Codebücher mit 16 Codebuchvektoren mit einer Dimension von 32. Die Ergebnisse der Testläufe sind in Tabelle 4.9 wiedergegeben. Diese sind schlechter als die Ergebnisse der BN/ESST-Akustik. Dies ist auf die Qualitätsverschlechterung der Aufnahmen durch die Reduzierung der Samplingrate und Filterung auf die benötigte Telefonqualität der Akustik zurückzuführen. Daraus kann gefolgert werden, dass für die betrachteten Verfahren eine „reine Laborakustik" als Basis verwendet werden sollte.

Verfahren	Luftfahrzeug	CFG	LM
MAM mit SCMS und zweifacher bereichskombinierter Sprach-Pause-Detektion	Helikopter	WFR 27,14% SFR 55,71%	WFR 24,21% SFR 55,71%
	Sportflugzeug	WFR 23,87% SFR 44,67%	WFR 9,71% SFR 24,67%
spektrale Subtraktion mit Störgeräusch-schätzung durch Minimum Statistik	Helikopter	WFR 33,25% SFR 62,86%	WFR 28,36% SFR 58,57%
	Sportflugzeug	WFR 23,05% SFR 45,13%	WFR 9,48% SFR 24,11%

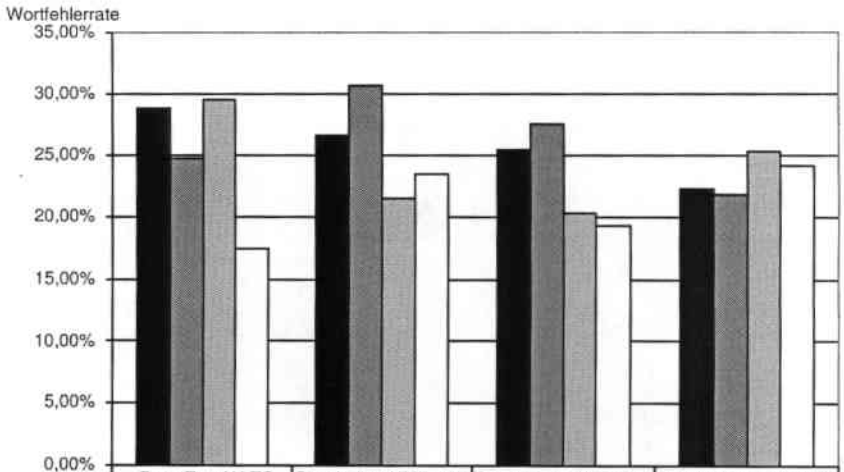
Tabelle 4.9: Erkennungsergebnisse mit der Switchboard-Akustik

Test des Spracherkennungssystems

Abschließend wurden Testläufe auf den Testdaten aus dem Helikopter, aufgenommen mit dem Bose- und dem Sennheiser-Headset, und aus dem Sportflugzeug, aufgenommen mit dem Sennheiser-Headset, durchgeführt. Dabei ist zu beachten, dass die Texte der Testdaten der mit dem Sennheiser-Headset gemachten Aufnahmen sich von den Texten der Entwicklungsdaten durch das wiederholte Sprechen der Anfragen nicht unterscheiden. Eine disjunkte Schnittmenge von Testtexten und Entwicklungstexten ist nur bei den Aufnahmen im Helikopter mit dem Bose-Headset gegeben. Aus diesem Grund wurde auch nur in diesem Fall das trigram-basierte Sprachmodell zum Vergleich herangezogen. In Abbildung 4.6 sind die Ergebnisse der einzelnen Verfahren angegeben. In diesen abschließenden Versuchen wurde ein lattice-rescoring verwendet, so dass das beste Ergebnis über verschiedene Parameterkombinationen der Parameter lz und lp ermittelt wurde. Aufgrund der Ergebnisse auf den Entwicklungsdaten wurde die modifizierte MAM für einen Einsatz in einem Spracherkennungssystem für Luftfahrzeuge vorgeschlagen. Diese Wahl wurde auf den Testdaten verifiziert.

Es ist deutlich zu sehen, dass die kontextfreie Grammatik bei unbekannten Daten wesentlich besser ist als das trigram-basierte Sprachmodell. Die modifizierte MAM zeigt sich in den verschiedenen Geräuschsituationen der spektralen Subtraktion überlegen. Die Aufnahmen aus dem Helikopter, die mit dem Sennheiser-Headset gemacht wurden, sind schon aufgrund der Störgeräuschanalyse schlechter zu beurteilen als die Aufnahmen mit dem Bose-Headset, weshalb auch die Wortfehlerrate deutlich höher ausfällt. Die Wortfehlerrate der MAM ist bei den ungestörten Sprachdaten höher als die der Standard-Vorverarbeitung. Dies ist auf die Geräuschreduktion durch die akustische Transformation zurückzuführen. Bessere Ergebnisse auf den Labordaten liefert die spektrale Subtraktion, deren Parameter auf den Entwicklungsdaten des Helikopters ermittelt wurden. Dadurch lässt sich auch die Verschlechterung der spektralen Subtraktion bei den Sportflugzeugdaten erklären. Der Schwellwert des Sprach-Pause-Detektors der MAM wurde ebenfalls auf den Helikopter-Entwicklungsdaten bestimmt. Deshalb ist es überraschend, dass die Wortfehlerrate bei den Testdaten aus dem Helikopter mit dem Bose-Headset aufgenommen sehr hoch, mit dem Sennheiser-Headset aufgenommen sehr niedrig ist.

In den zahlreichen Experimenten wurden verschiedene signalbasierende Verfahren auf ihren Einsatz in einem Spracherkennungssystem für Luftfahrzeuge getestet. Wei-



	Bose Typ AH-TC Helikopter MD-520N	Sennheiser HME 100 Helikopter MD-520N	Sennheiser HME 100 Sportflugzeug TB 20	Sennheiser HME 100 Labor
Standard-Vorverarbeitung (Baseline)	28,86%	26,67%	25,47%	22,33%
Spektrale Subtraktion	24,83%	30,69%	27,64%	21,85%
MAM	29,53%	21,56%	20,41%	25,39%
modifizierte MAM	17,45%	23,51%	19,30%	24,23%

Abbildung 4.6: Ergebnisse auf den Testdaten

terhin wurden verschiedene Akustiken sowie zwei Sprachmodelle evaluiert. Aufgrund der durchgeführten Experimente und gemachten Erläuterungen wurde das beste Verfahren bestimmt. Die modifizierte MAM ist in einem Spracherkennungssystem mit kontextfreier Grammatik als Sprachmodell und mit der BN/ESST-„Laborakustik" für eine Spracherkennung im Cockpit von Luftfahrzeugen am besten geeignet. Ein Geschwindigkeitstest dieses Systems ergab einen RTF von 0,6 auf einem Computer mit einem Pentium-IV-Prozessor mit 2,66 GHz.

5 Zusammenfassung und Ausblick

Spracherkennung in Umgebungen mit starken Störgeräuschen stellt noch immer eine Herausforderung dar. Spracherkennung in Luftfahrzeugen mit sogenannten „Laborakustiken“ ist eine enorme Herausforderung. Zahlreiche Vorschläge zur Geräuschreduktion sind vorhanden, Verfahren gibt es viele. Motivation für diese Arbeit war eine Erhöhung der Flugsicherheit durch den Einsatz von Spracherkennungstechnologie im Cockpit. Manche Arbeitsabläufe könnten erleichtert und Einstellungsfehler vermieden werden. Dadurch wird der Pilot entlastet und kann sich auf seine eigentliche Aufgabe – das Führen des Luftfahrzeugs – konzentrieren.

In dieser Arbeit wurden zur Geräuschreduktion die spektrale Subtraktion mit entsprechender Störgeräuschschätzung, die modellkombinationsbasierte akustische Transformation (MAM) und die sprachbasierte cepstrale Mittelwertsubtraktion ausgewählt und auf einen Einsatz im Luftfahrzeug anhand weniger Sprachaufnahmen weniger Sprecher aus einem Helikopter und einem Sportflugzeug untersucht. Die bereichskombinierte Sprach-Pause-Detektion wurde entwickelt und mit ihr eine Verbesserung der Erkennungsraten erzielt. Diese sowie die sprachbasierte cepstrale Mittelwertsubtraktion wurden in allen Verfahren verwendet. Einige betrachtete Verfahren wurden in den abschließenden Tests auf den Testdaten benutzt. Bei den Testergebnissen, die in Abbildung 5.1 angegeben sind, ist allerdings zu beachten, dass die Parameter der Störgeräuschschätzung der spektralen Subtraktion auf den Entwicklungsdaten des Helikopters bestimmt wurden. Weiterhin ist zu beachten, dass nur für die mit dem Bose-Headset gemachten Aufnahmen im Helikopter die Forderung einer disjunkten Schnittmenge von Texten der Test- und Entwicklungsdaten gegeben ist. Deshalb wurde nur in diesem Fall das trigram-basierte Sprachmodell zu einem fairen Vergleich herangezogen.

Die modifizierte MAM kann als bestes Verfahren in einem Spracherkennungssystem mit BN/ESST-“Labor“-Akustik in Luftfahrzeugen eingesetzt werden. Allerdings sollten in Zukunft weitere (modellbasierte) Verfahren zur Geräuschreduktion betrachtet werden. Weiterhin muss das verwendete Sprachmodell durch mehr Daten verbessert werden. Eine flugphasenabhängige Einschränkung des Sprachmodells wäre denkbar. Die Kombination mit einer visuellen Erkennung der Lippenbewegungen könnte eine

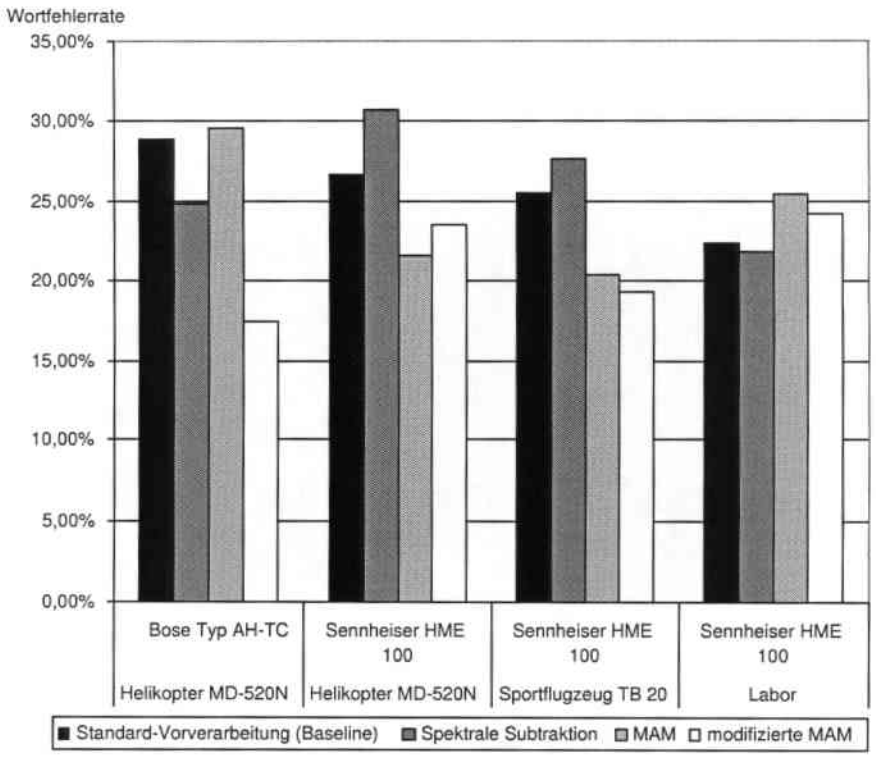


Abbildung 5.1: Ergebnisse auf den Testdaten

weitere Verbesserung der Erkennungsraten bewirken. Allerdings ist dazu eine umfangreiche und aufwändige Datensammlung notwendig.

Da davon auszugehen ist, dass das System während eines Fluges der Allgemeinen Luftfahrt nur von einem Pilot bedient wird, kann eine unüberwachte Sprecheradaption durch eine MLLR[1] erfolgen. Ist der Pilot bereit, das System durch das Sprechen von ein paar vorgegebenen Sätzen zu trainieren, so kann in diesem Fall eine überwachte Sprecheradaption erfolgen. Verwenden Pilot und Copilot gemeinsam das System, so sollte zusätzlich zur Sprachaufnahme über das Bordkommunikationsgerät auch die Herkunft der Sprache ermittelt werden. So kann wiederum eine unüberwachte Adaption auf den jeweiligen Sprecher durchgeführt werden.

In Zukunft wären unter Einsatz der in dieser Arbeit betrachteten Technologien folgende Anwendungen möglich:

- Ein neuartiges Flugplanungs- und durchführungsprogramm, das eine Flugplanung durch Spracheingabe mit einem Dialogsystem auf einem Notebook in einem „Labor" ermöglicht. Dieses Notebook sollte an die Systeme im Flugzeug angekoppelt werden können und dadurch den kompletten Flugverlauf überwachen und gegebenenfalls Änderungen durch eine Spracheingabe des Piloten vornehmen. Als Ausgabe sollte im Flugzeug nicht das Display des Notebooks und die Sprachausgabe, sondern ein im Panel des Cockpit eingebautes Multifunktionsdisplay (MFD) verwendet werden.

[1]Maximum Likelihood Linear Regression

- Die Systeme im Flugzeug könnten die automatischen Start- und Landeinformationen an Flugplätzen (ATIS[2]) über ein Funkgerät „abhören" und auf Wunsch des Piloten diese Informationen auf einem Multifunktionsdisplay (MFD) darstellen. Weiterhin sollte natürlich auch der ganze Funkverkehr von einem System „mitgehört" werden. Dann kann der Pilot ihm entgangene Funksprüche auf einem Display im Cockpit nachlesen. Dies entlastet den Funkverkehr und den Pilot und führt zu einer Erhöhung der Flugsicherheit.

- Unkontrollierte Flugplätze könnten automatisiert und damit 24 Stunden am Tag betrieben werden. Durch eine entsprechendes Spracherkennungssystem mit einem angeschlossenen Dialogsystem könnten dem Piloten über Funk alle wesentlichen Informationen auf Anfrage zur Verfügung gestellt werden. Weiterhin kann es per Funk zum Beispiel zum Anschalten der Landebahnbeleuchtung aufgefordert werden.

[2]Automatic Terminal Information System

A Datensammlung

Im Rahmen dieser Arbeit war eine Datensammlung notwendig. Diese gliederte sich in zwei Teile. Zuerst wurde eine Umfrage durchgeführt, deren Ergebnisse die Erstellung der Sprachmodelle vereinfachen sollte. Danach wurden Sprachaufnahmen in einem Sportflugzeug und einem Helikopter sowie im Labor aufgenommen. Im Jahre 1999 wurden am Institut für Logik, Komplexität und Deduktionssysteme der Universität Karlsruhe (TH) bereits Sprachaufnahmen in einem Hubschrauber des Typs MD-520N gemacht. Diese Aufnahmen wurden bei der vorliegenden Arbeit ebenfalls verwendet. Die Umfrage und die zur Datensammlung verwendeten Geräte werden im Folgenden beschrieben.

A.1 Umfrage

Im Vorfeld dieser Arbeit wurde eine Umfrage zur Erlangung von spontanen Anfragen an ein durch Spracherkennung gesteuertes GPS durchgeführt. Eine schriftliche Anfrage wurde an die akademische Fliegergruppe der Universität Karlsruhe (TH), an die Motorfluggemeinschaft Speyer sowie folgende Newsgroups gestellt:

- de.rec.luftfahrt
- alt.aviation.safety
- eunet.aviation
- rec.aviation.ifr
- rec.aviation.misc
- rec.aviation.piloting
- rec.aviation.products
- rec.aviation.rotorcraft
- rec.aviation.student

- uk.rec.aviation

Auf diese Umfrage kamen leider nur 14 Antworten zurück. Die Anfragen an das GPS wurden von allen Befragten ähnlich gestellt. Es wurde dabei immer wieder die Sprache im Funkverkehr genannt, die kommandobasiert nach Regeln aufgebaut ist und eine effektive Kommunikation durch kurze „Sätze" ermöglicht.

Als Beispiele seien hier genannt:

- show wind correction angle
- report time to < *waypoint* >
- report groundspeed
- show moving map
- show current track
- make a new route from < *town* > to < *town* >
- show map / HSI / CDI
- zoom in / out
- display nearest weather
- show info about < *NDB / VOR / airport* >
- report vertical speed to cross < *waypoint* > at < *altitude* >

Die Anworten wurden als Grundlage für den Aufbau der Sprachmodelle, beschrieben in Kapitel 4.3, verwendet.

A.2 Sprachaufnahmen

Auf einem Flug von Karlsruhe nach Flensburg wurden im September 1999 Sprachaufnahmen in einem Hubschrauber des Typs MD-520N gemacht. Ein DAT-Rekorder wurde hierfür mit einem Adapter direkt an das Bordkommunikationssystem angeschlossen. Als Mikrofon wurde das eingebaute Mikrofon der Headsets von Bose, Typ AH-TC, verwendet. Aus der Bedienungsanleitung sowie sonstiger Publikationen von Bose geht der Übertragungsbereich des Headset nicht hervor. Ein Telefongespräch mit einem Bose-Ingenieur brachte Klarheit. Nach dessen Angaben ist der Übertragungsbereich des Elektret-Mikrofons 300 bis etwa 3000 - 3500 Hz, was in etwa dem Übertragungsbereich eines analogen ITU-Standard-Telefonkanals (300 - 3400 Hz) entspricht.

Auf einem weiteren Flug mit dem Helikopter wurden mit einem Standard-Nahbesprechungsmikrofon der Firma Sennheiser HD-440-6 Sprachaufnahmen gemacht, die in dieser Arbeit nur zur Betrachtung der Geräuschsituation verwendet wurden.

Die Aufnahmen im Sportflugzeug wurden mit Headsets des Typs Sennheiser HME 100 im Rahmen dieser Arbeit durchgeführt. In diese Headsets ist das Mikrofon MKE

45-1 P/N eingebaut. Dieses Mikrofon besitzt einen Vorverstärker und arbeitet geräuschkompensierend. Der Übertragungsbereich wird mit 300 - 5000 Hz angegeben. Zur Aufnahme wurden zwei HME 100 an jeweils eine Intercom des Typs SL 400 angeschlossen. Diese Bordkommunikationsgeräte haben einen Ausgang für einen Rekorder. Dieser Ausgang wurde für die Aufnahme mit einem DAT-Rekorder benutzt. Die Aufnahmen erfolgten auf zwei Kanälen, um mit dem einen Mikrofon nur die Geräusche und mit dem anderen Geräusche und Sprache aufzunehmen.

Die Sprachdaten der Aufnahmen aus dem Hubschrauber und aus dem Sportflugzeug wurden mit einem DAT-Rekorder Sony TDC-8 mit einer Abtastfrequenz von 48 kHz digital aufgezeichnet. Danach wurden die Daten über eine spezielle Hardware-Karte auf einen PC übertragen und mit Hilfe der Software CoolEdit[1] von Hand segmentiert und mit einer Auflösung von 16 bit und einer Abtastfrequenz von 48 kHz im WAV-Format abgespeichert. Zur Verwendung im Spracherkenner wurden die Daten danach automatisch auf eine Abtastfrequenz von 16 kHz im RAW-Format transformiert.

Zur besseren Vergleichbarkeit der erhaltenen Ergebnisse wurden im Rahmen dieser Arbeit nochmals Sprachaufnahmen im Helikopter gemacht, allerdings nicht mit dem Bose Headset, sondern mit der gleichen Hardware, die auch bei den Aufnahmen im Sportflugzeug verwendet wurde. Die Aufnahmen wurden dabei nicht mit einem DAT-Rekorder aufgezeichnet, sondern mit Hilfe eines Notebooks mit einer Abtastfrequenz von 16 kHz und einer Auflösung von 16 bit im RAW-Format gespeichert.

Da manche Systeme im Luftfahrzeug vor dem Anlassen der Motoren eingestellt werden, sollte ein Spracherkennungssystem auch „Labor"-Sprache erkennen können. Deshalb wurden mit dem Sennheiser-Headset HME 100, der Intercom SL 400 und einem Notebook die im Anhang A.5 aufgeführten Anfragen an ein Flugnavigationssystem im Labor aufgenommen. Die Daten wurden ebenfalls mit einer Abtastfrequenz von 16 kHz und einer Auflösung von 16 bit im RAW-Format gespeichert.

A.3 Luftfahrzeuge

Ein Hubschrauber des Typs MD-520N und ein Sportflugzeug des Typs Socata TB20 Trinidad GT standen für die Datensammlung zur Verfügung. Die technischen Daten beider Luftfahrzeuge werden im Folgenden kurz dargestellt.

Helikopter MD-520N

Der MD-520N fällt in die Klasse der turbinenangetriebenen Mehrzweck-Helikopter. Es war der erste Helikopter mit NOTAR[2]-System. Es verwendet zur Steuerung des Helikopters um die Hochachse anstatt des üblichen Heckrotors ein „Gebläse", welches einen zur Steuerung ausreichenden steuerbaren Luftstrom erzeugt. Es verbessert das Flugverhalten und die Steuerung und erhöhte die Sicherheit am Boden. Der MD-520N ist ein vier- oder fünfsitziger Helikopter. Angetrieben wird der Hauptrotor sowie das NOTAR-System durch eine Turbine des Typs Allison 250-C20R. Mit einer Tankfüllung kann der Helikopter maximal eine Strecke von 389 Kilometern bewältigen. Seine Höchstgeschwindigkeit liegt bei 281 km/h.

[1] Die verwendete Software CoolEdit war ein eingetragenes Warenzeichen der Firma Syntrillium Software. Im Mai 2003 wurde Syntrillium Software von Adobe übernommen. Adobe vertreibt das CoolEdit-Paket nun unter dem Namen Adobe Audition.

[2] NO TAil Rotor

Abbildung A.1: Helikopter MD-520N [Quelle: www.airliners.net]

Abbildung A.2: Innenansicht eines Helikopter MD-520N

Sportflugzeug Socata TB 20 Trinidad GT

Die TB 20 fällt in die Klasse der High-Performance-Flugzeuge und Tiefdecker. Sie ist ein viersitziges Reiseflugzeug mit Verstellpropeller und Einziehfahrwerk. Die für die Sprachaufnahmen verwendete TB 20 ist für Instrumentenflugbetrieb ausgerüstet und besitzt neben einem Wetterradar auch eine Anti-Ice-Ausrüstung. Die Tanks in den Tragflächen verfügen über 380 Liter Fassungsvermögen. Damit hat die TB 20 mit einem 250 PS Motor (Lycoming IO-540) eine maximale Flugzeit von circa 4,5 Stunden.

Abbildung A.3: Sportflugzeug Socata TB 20 Trinidad GT [Quelle: www.airliners.net]

Abbildung A.4: Innenansicht eines Sportflugzeug Socata TB 20 Trinidad GT [Quelle: www.airliners.net]

A.4 Störgeräusche

Im Kapitel 2.1 wurden die Störgeräusche der Luftfahrzeuge analysiert. Die verwendeten Spektren und deren Varianz werden in diesem Abschnitt abgebildet. Die entsprechenden Werte werden nochmals mit einer logarithmierten Y-Achse dargestellt, da die Störgeräuschwerte der hohen Frequenzbänder aus der unlogarithmierten Darstellung nicht hervorgehen.

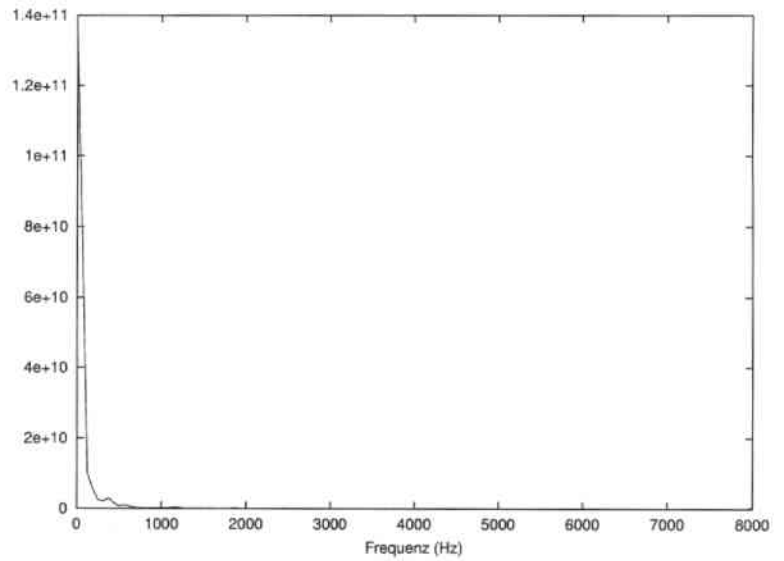


Abbildung A.5: Durchschnittliches Störgeräusch aus dem Helikopter, aufgenommen mit dem Mikrofon Sennheiser HD-440-6

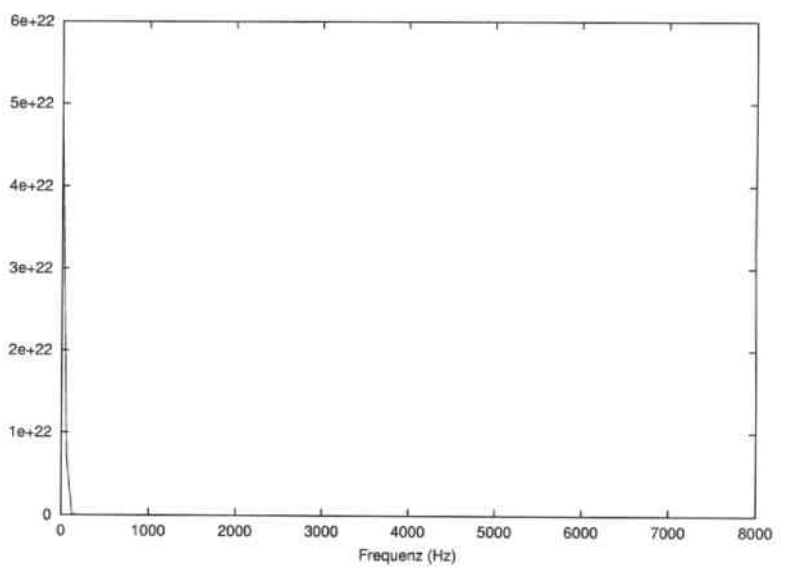


Abbildung A.6: Varianz des Störgeräuschs aus dem Helikopter, aufgenommen mit dem Mikrofon Sennheiser HD-440-6

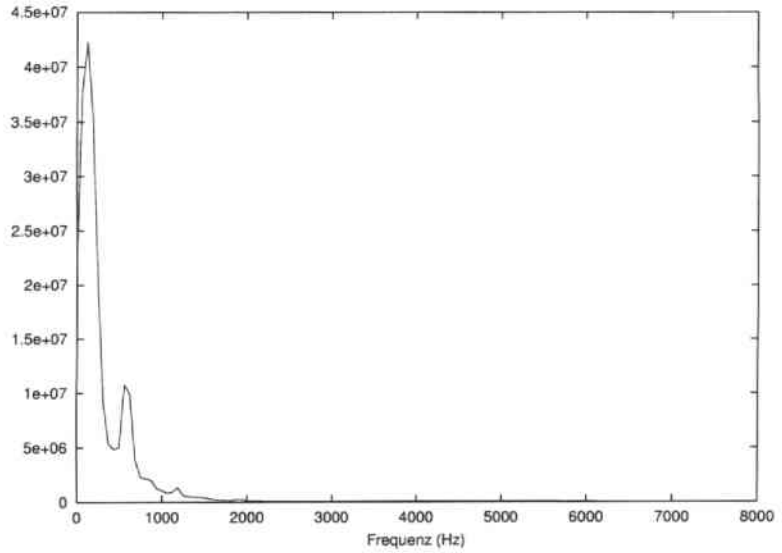


Abbildung A.7: Durchschnittliches Störgeräusch aus dem Helikopter, aufgenommen mit dem Mikrofon des Headsets Bose Typ AH-TC

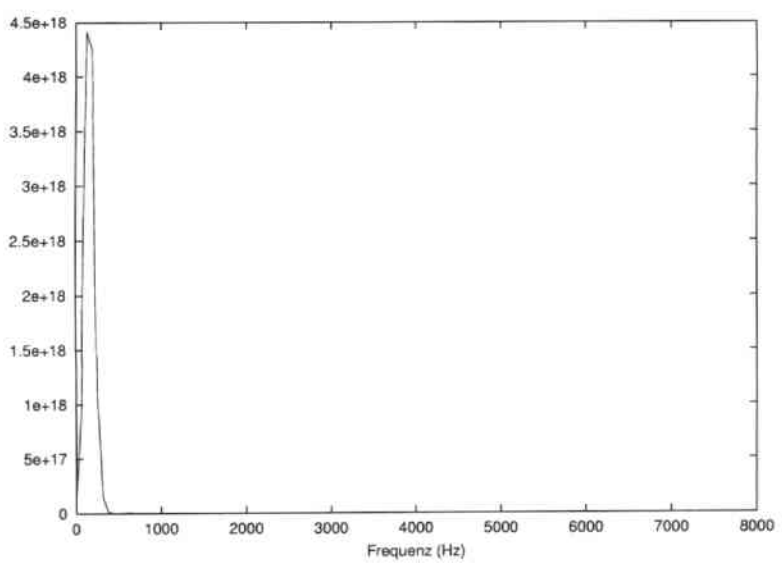


Abbildung A.8: Varianz des Störgeräuschs aus dem Helikopter, aufgenommen mit dem Mikrofon des Headsets Bose Typ AH-TC

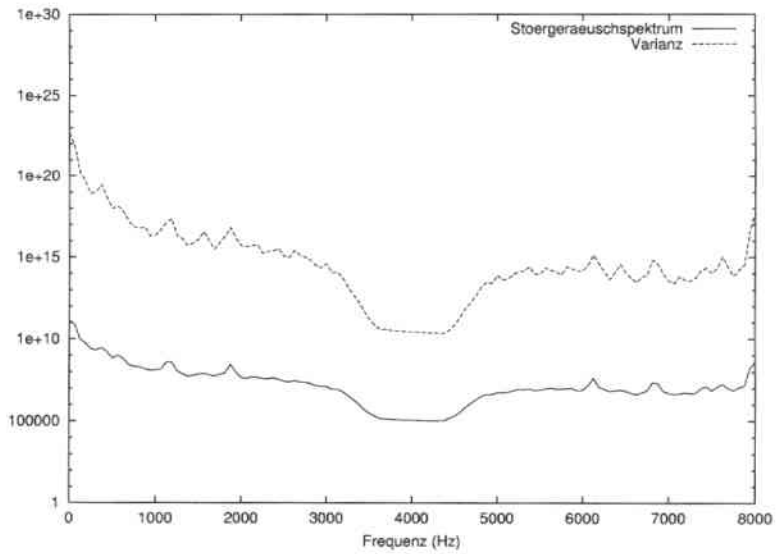


Abbildung A.13: Durchschnittliches Störgeräusch und Varianz des Störgeräuschs aus dem Helikopter, aufgenommen mit dem Mikrofon Sennheiser HD-440-6 (Abbildung mit logarithmierter Y-Achse)

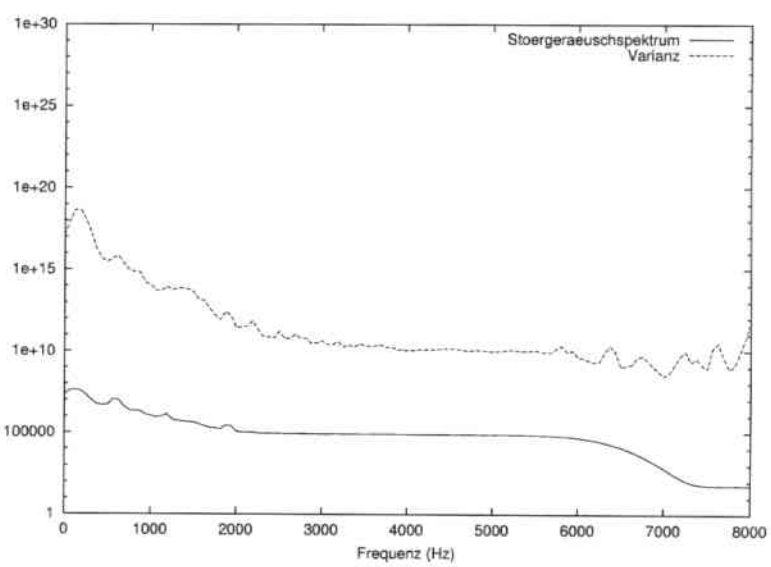


Abbildung A.14: Durchschnittliches Störgeräusch und Varianz des Störgeräuschs aus dem Helikopter, aufgenommen mit dem Mikrofon des Headsets Bose Typ AH-TC (Abbildung mit logarithmierter Y-Achse)

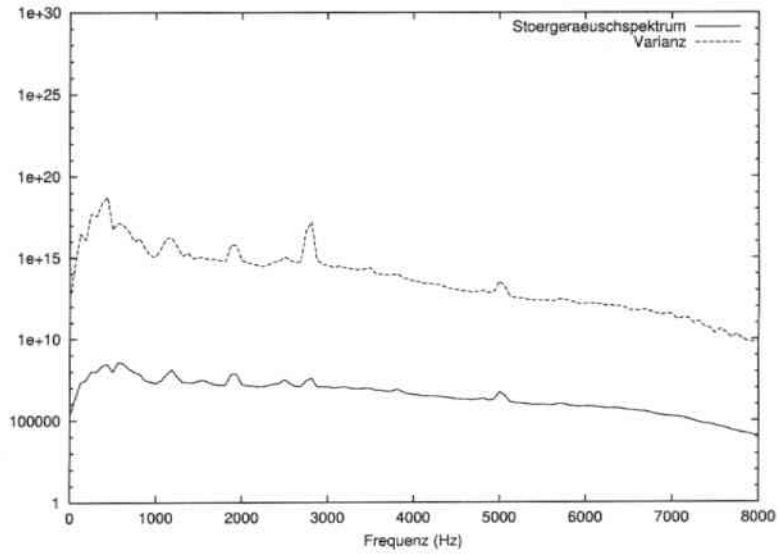


Abbildung A.15: Durchschnittliches Störgeräusch und Varianz des Störgeräuschs aus dem Helikopter, aufgenommen mit dem Mikrofon des Headsets Sennheiser HME 100 (Abbildung mit logarithmierter Y-Achse)

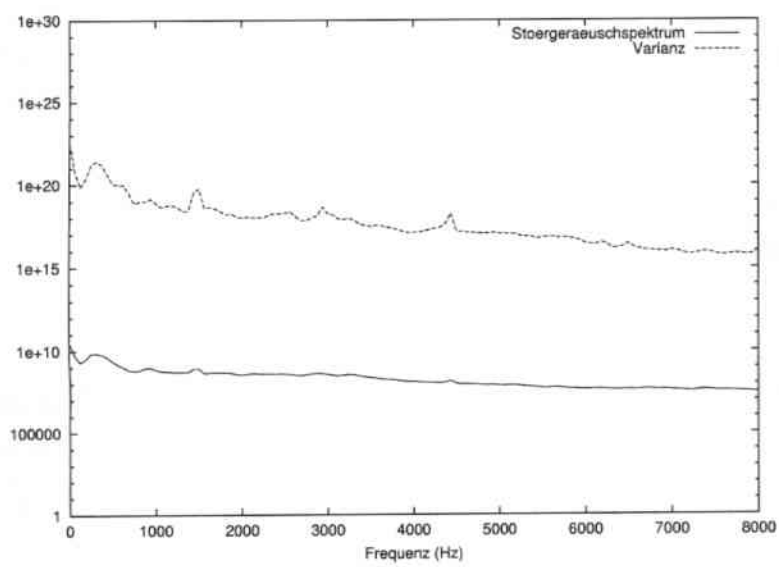


Abbildung A.16: Durchschnittliches Störgeräusch und Varianz des Störgeräuschs aus dem Sportflugzeug, aufgenommen mit dem Mikrofon des Headsets Sennheiser HME 100 (Abbildung mit logarithmierter Y-Achse)

A.5 Anfragen an ein Flugnavigationssystem

Mit den Umfrageergebnissen als Ausgangspunkt wurde im Rahmen dieser Arbeit ein kleiner Katalog von möglichst spontanen Anfragen an ein Flugnavigationssystem erstellt. Die Erstellung erfolgte zum Teil vor den Sprachaufnahmen im Cockpit, der andere Teil wurde spontan bei den Aufnahmen gesprochen und danach verschriftet. Die verwendeten 214 Anfragen werden auf den folgenden Seiten alphabetisch sortiert angegeben.

- activate ils approach runway two seven
- activate ils approach runway zero nine
- activate the inverted route
- activate the loaded route
- activate the route
- activate the route from karlsruhe baden to echo delta november yankee
- add the actual position to the activated route
- add the approach route runway zero nine to the activated route
- add the departure route runway two seven to the activated route
- after landing set com two to karlsruhe ground
- after landing switch com one to karlsruhe ground
- alert me five minutes before reaching karlsruhe baden
- alert me two minutes before reaching
- change the display to map
- check mannheim atis at frequency one three six point five five
- check nearest atis frequency
- check the nearest weather frequency
- delete the next waypoint from the actual route
- dial frankfurt radar
- dial in the frequency for heidelberg
- display the actual track
- display the actual weather
- display the ete
- do not warn airspace charlie

- don't warn airspace charlie
- don't warn any airspace
- enter a new waypoint at the current position
- find a route from echo delta sierra bravo to echo delta foxtrott mike
- find a route from flensburg to karlsruhe
- find a route from flensburg to stuttgart
- find a route to karlsruhe
- find a short route from munich to heidelberg
- find a taxi route to apron one
- from current position fly to karlsruhe baden
- from current position fly to the nearest airport
- from current position fly to the next airport
- give alert when reaching a controlzone
- give information about echo delta kilo alpha
- give me info about the mannheim departure route
- give me information about the mannheim departure route
- give me the actual flap position
- give me the actual position of the flaps
- give me the approach chart of heidelberg
- give me the standard arrival routes from echo delta sierra bravo
- give me the star of echo delta sierra bravo
- give me the visual approach chart of echo delta romeo yankee
- give me the visual approach chart of heidelberg
- give me the visual approach chart of mannheim
- go to echo delta november yankee via entry point romeo
- go to karlsruhe vor
- go to speyer ndb
- go to the moving map
- go to the nearest airport

- go to this new waypoint
- go to this waypoint
- help i'm lost
- i'd like to enter a new route
- i'd like to enter a new waypoint
- i'd like to taxi to the gat via alpha and delta
- i'm approaching echo delta sierra bravo via entry point romeo
- i'm lost please help me
- insert a new waypoint
- insert my actual position into the current route
- integrate the nearest waypoint into the activated route
- invert the current route
- load the route from stuttgart to frankfurt
- load the route to echo delta delta fox
- make a new user waypoint
- new communication frequency one two one decimal six
- on which track do we fly
- open airfield echo delta tango uniform
- please check atis at frequency one three six decimal five five
- please check atis frequency one three six decimal five five
- please check frankfurt atis
- please check mannheim atis at frequency one three six point five five
- please display the ete
- please give me the actual weather information
- please give me the estimated time of arrival
- please load the route from echo delta foxtrott mike to karlsruhe baden
- please load the route from karlsruhe to echo delta romeo yankee
- please load the route to echo delta delta fox
- please load the route to stuttgart

- please make a new route
- please report the distance to berlin
- please report the time to karlsruhe baden
- please show me the approach chart of echo delta romeo yankee
- please show me the approach chart of echo delta sierra bravo
- please show me the course
- please show me the data fields
- please show me the distance to frankfurt airport
- please show me the nearest airports
- please show me the nearest intersection
- please show me the nearest vor
- please show me the two nearest airports
- please show the approach chart
- please show the frequency of mannheim tower
- please show the hsi page
- please show the wind correction angle
- please squawk twenty two when reaching altitude five thousand feet
- please tell me the vertical speed to cross echo delta sierra bravo at four thousand five hundred feet
- please tell me the vertical speed to cross karlsruhe vor at two thousand feet
- please warn at airspace delta and charlie
- report the airspace in front of us
- report the distance to echo delta delta mike
- report the distance to the next waypoint
- report the groundspeed
- report the vertical speed to cross karlsruhe vor at flightlevel six five
- report time to echo delta sierra bravo
- report time to the next waypoint
- set com one frequency one two three point five five
- set com one frequency one two three point two five

- set com one frequency one two zero decimal four five
- set com one to karlsruhe ground
- set com two karlsruhe ground
- set course to karlsruhe baden
- set nav one ils frequency runway three five
- set nav one to ils runway two seven
- set nav two frequency one zero eight decimal four five
- set nav two frequency one zero eight decimal six zero
- set nav two ils frequency runway two seven
- set nav two one zero eight decimal six zero
- set nav two to karlsruhe vor
- set navigation frequency to ils
- set navigation to karlsruhe vor
- set the course to echo delta fox mike
- set the course to echo delta sierra bravo
- set the course to karlsruhe vor
- set the navigation frequency to ils
- set transponder mode charlie zero zero two one
- set transponder to emergency
- show current track
- show desired track
- show me apron one
- show me available routes
- show me information about karlsruhe vor
- show me my actual position
- show me the actual groundspeed
- show me the altitude
- show me the approach chart of echo delta foxtrott mike
- show me the apron

- show me the arrival route of munich
- show me the available routes
- show me the bearing
- show me the current flap position
- show me the departure route
- show me the departure route from echo delta foxtrott mike via point whisky
- show me the departure route from echo delta foxtrott mike via whisky
- show me the departure route of echo delta foxtrott mike
- show me the downwind runway two seven
- show me the estimated time elapsed
- show me the flap position
- show me the heading to ried vor
- show me the heading to the ried vor
- show me the ifr chart of karlsruhe baden
- show me the mannheim departure route
- show me the nearest airport
- show me the qdm to ried vor
- show me the qdm to the ried vor
- show me the right handed traffic circuit runway two seven
- show me the right handed traffic circuit runway zero nine of mannheim
- show me the right handed traffic pattern
- show me the sierra departure route runway two seven
- show me the speyer traffic circuit runway three five
- show me the taxiway alpha and delta
- show me the taxiway alpha delta and sierra
- show me the taxiways after landing
- show me the traffic circuit
- show me the traffic circuit of echo delta romeo yankee
- show me the traffic circuit of echo delta sierra bravo

- show me the traffic circuit runway one seven of speyer
- show me the traffic circuit runway two seven
- show me the traffic circuit runway zero two
- show me the traffic pattern runway two seven
- show me the whisky departure route
- show me the whisky departure route runway two seven
- show me the whisky departure route runway zero nine
- show the correction angle
- show the crosswind
- show the estimated time elapsed
- show the eta
- show the frequency of echo delta romeo yankee
- show the gps
- show the gps message
- show the map
- show the message
- show the moving map
- show the nearest ndb
- show the nearest weather station
- show the preflight checklist
- squawk communication failure
- squawk ident
- squawk three seven zero one
- squawk twenty one
- tell me more about speyer ndb
- tell me the actual wind
- tell me the airspace we are flying in
- tell me the altitude
- tell me the distance to mannheim airport

- tell me the estimated elapsed time
- tell me the estimated time elapsed
- tell me the frequency of frankfurt radar
- tell me the name of the town five miles ahead
- tell me the name of the town in front
- we are approaching runway two six
- which altitude do we actual have
- which groundspeed do we have
- which heading do we fly at the moment
- which track do we fly
- zoom in
- zoom in five steps
- zoom in two steps
- zoom out
- zoom out four steps
- zoom out two steps

Literatur

[Ande98] Timothy R. Anderson. Applications Of Speech-Based Control. In *RTO Lecture Series 215 - Alternative Control Technologies: Human Factors Issues*, Neuilly-Sur-Seine Cedex, 1998. North Atlantic Treaty Organization - Research and Technology Organization, RTO-EN-3.

[Beek01] Douglas W. Beeks. Speech Recognition and Synthesis. In Cary R. Spitzer (Hrsg.), *The Avionics Handbook*, Kapitel 8. CRC Press, 2001.

[Boll79] S. Boll. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band 27, No. 2, 1979, S. 113–120.

[ChANK01] Y. Cho, K. Al-Naimi und A. Kondoz. Improved voice activity detection based on a smoothed statistical likelihood ratio. In *Proc. Int. Conf. Acoust., Speech and Signal Processing*, 2001.

[dVBo96] J. de Veth und L. Boves. Comparison of Channel Normalisation Techniques for Automatic Speech Recognition Over the Phone. In *Proc. ICSLP '96*, Band 4, Philadelphia, PA, 1996. S. 2332–2335.

[Euro01] Team Eurofighter. DVI - Making the Difference. In *Team Eurofighter - Issue 1-2001.* Eurofighter GmbH, München, 2001.

[FGHK+97] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries und M. Westphal. The Karlsruhe-Verbmobil Speech Recognition Engine. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-97*, Munich, 1997.

[FSSM+03] C. Fügen, S. Stüker, H. Soltau, F. Metze und T. Schultz. Efficient Handling of Multilingual Language Models. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU 2003)*, St. Thomas, U.S. Virgin Islands, 2003.

[Gale95] M.J.F. Gales. *Model-Based Techniques for Noise Robust Speech Recognition.* Cambridge University, PhD-Thesis, Cambridge. 1995.

[Gale98] M.J.F. Gales. Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. In *Computer Speech and Language*, Band 12, 1998, S. 75–98.

[Gerl96] Marc Gerlach. *Schnittstellengestaltung für ein Cockpitassistenzsystem unter besonderer Berücksichtigung von Spracheingabe.* VDI Verlag, Düsseldorf. 1996.

[HaMa93a] J. Haigh und J. Mason. Robust voice activity detection using cepstral features. In *IEEE TENCON*, China, 1993. S. 321–324.

[HaMa93b] J. Haigh und J. Mason. A voice activity detector based on cepstral analysis. In *Proc. European Conf. Speech Communication and Technology*, Band 2, 1993, S. 1103–1106.

[Hayk96] Simon Haykin. *Adaptive Filter Theory.* Prentice Hall, Upper Saddle River, New Jersey. 1996.

[IlKa99] Georgie Iliev und Nikola Kasabov. Adaptive Filtering with Averaging in Noise Cancellation for Voice and Speech Recognition. In *Future Directions for Intelligent Systems and Information Sciences*, Dunedin, New Zealand, 1999. International Conference on Neural Information Processing.

[KaKr89] K. D. Kammeyer und K. Kroschel. *Digitale Signalverarbeitung, Filterung und Spektralanalyse.* Teubner Studienbücher Elektrotechnik, Stuttgart. 1989.

[KaKr98] K.D. Kammeyer und K. Kroschel. *Digitale Signalverarbeitung, Filterung und Spektralanalyse mit MATLAB-Übungen.* Teubner Verlag, Stuttgart. 1998.

[KeMo99] C. Kermorvant und A. Morris. A comparison of two strategies for ASR in additive noise: Missing Data and Spectral Subtraction. In *Proc. Eurospeech'99*, 1999, S. 2891–2844.

[Kerm99] C. Kermorvant. A comparison of noise reduction techniques for robust speech recognition. In *IDIAP-RR 10*, 1999.

[Kros96] Kristian Kroschel. *Statistische Nachrichtentheorie.* Springer Verlag, Berlin, Heidelberg. 1996.

[KSYC+00] Jae Won Kim, Min Sik Seo, Byung Sik Yoon, Song In Choi und Young Gap You. A Voice Activity Detection Algorithm for Wireless Communication Systems with Dynamically Varying Background Noise. In *IEICE Trans. Commun.*, Band E83-B, 2000, S. 414–418.

[Lege98] Alain Leger. Synthesis- and expected benefits Analysis. In *RTO Lecture Series 215 - Alternative Control Technologies: Human Factors Issues*, Neuilly-Sur-Seine Cedex, 1998. North Atlantic Treaty Organization - Research and Technology Organization, RTO-EN-3.

[Leis99] John Leis. *Adaptive Filter Lecture Notes and Examples.* University of Southern Queensland, Toowoomba, Australia. 1999.

[Loga98] B.T. Logan. *Adaptive Model-Based Speech Enhancement.* Cambridge University, PhD-Thesis. 1998.

[Mart93] Rainer Martin. An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals. Berlin, 1993. EUROSPEECH, S. 1093–1096.

[McBa95] Alan V. McCree und Thomas P. Barnwell. A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding. Band 3 No. 4. IEEE Transactions an Speech and Audio Processing, 1995.

[MeSi97] J. Meyer und K. U. Simmer. Multi-Channel Speech Enhancement in a Car Environment using Wiener Filtering and Spectral Subtraction. In *Proc. ICASSP '97*, Munich, Germany, 1997. S. 1167–1170.

[MeSK97] Joerg Meyer, Klaus Uwe Simmer und Karl Dirk Kammeyer. Comparison Of One- And Two-Channel Noise-Estimation Techniques. In *Proc. 5th International Workshop on Acoustic Echo and Noise Control (IWAENC-97)*, Band 1, London, 1997. S. 17–20.

[NLR02] NLR. *FalconEar - Fighter Cockpit Voice Control.* National Aerospace Laboratory NLR, Amsterdam, Netherland. 2002.

[NoMa] Jan Novotny und Lukas Machacek. *Noise Reduction Applied in Real Time Speech Recognition System.* Department of Circuit Theory, Faculty of Electrical Engineering, University of Prague, Prague.

[Offe97] H. Offerman. First results from operating the Dutch national simulation facility NSF. In *International Training Equipment Conference (ITEC)*, Lausanne, Switzerland, 1997.

[PoSU95] P. Pollak, P. Sovka und J. Uhlir. Cepstral Speech/Pause Detectors. In *Proceedings of 1995 IEEE Workshop on Nonlinear Signal and Image Processing*, 1995, S. 388–391.

[RiSG97] Klaus Ries, Bernhard Suhm und Petra Geutner. Language Modeling in JANUS. Pittsburgh, 1997. School of Computer Science, Carnegie Mellon University, http://isl.ira.uka.de/~jrtk/doc.LM/janus-lm.doku.html.

[Rood98a] G. M. Rood. Human Factors Issues for the Integration of Alternative Control Technologies. In *RTO Lecture Series 215 - Alternative Control Technologies: Human Factors Issues*, Neuilly-Sur-Seine Cedex, 1998. North Atlantic Treaty Organization - Research and Technology Organization, RTO-EN-3.

[Rood98b] G. M. Rood. Operational Rationale and Related Issues for Alternative Control Technologies. In *RTO Lecture Series 215 - Alternative Control Technologies: Human Factors Issues*, Neuilly-Sur-Seine Cedex, 1998. North Atlantic Treaty Organization - Research and Technology Organization, RTO-EN-3.

[Schu95] E. G. Schukat-Talamazzini. *Automatische Spracherkennung - Statistische Verfahren der Musteranalyse.* Vieweg Verlag, Braunschweig, Wiesbaden. 1995.

[Schu01] Tanja Schultz. *Multilinguale Spracherkennung.* Shaker Verlag, Aachen. Dissertation an der Universität Karlsruhe (TH), 2001.

[Sing01] Aarti Singh. *Adaptive Noise Cancellation.* Central Electronics Engineering Research Institute, University of Delhi, Delhi, India. 2001.

[SMFW01] H. Soltau, F. Metze, C. Fügen und A. Waibel. A One Pass-Decoder based an Polymorphic Linguistic Context Assignment. In *Proc. of the Automation Speech Recognition and Understanding Workshop, ASRU-2001*, Madonna di Campiglio, Trento, Italy, 2001.

[SMSG+01] Abhijeet Sangwan, Chiranth M.C., Rahul Sah, Vishal Gaurav und R. Venkatesha Prasad. Voice Activity Detection for VoIP — Time and Frequency Domain Solution. In *10th Annual Symposium.* IEEE Bangalore Section, 2001.

[SoPo95] P. Sovka und P. Pollak. The Study of Speech/Pause Detectors for Speech Enhancement Methods. In *Proceedings of the 4th European Conference on Speech Communication and Technology*, 1995, S. 1575–1578.

[Stad99] Jan Stadermann. *Sprach/Pause-Detektion in der automatischen Spracherkennung.* Diplomarbeit an den Philips Forschungslaboratorien Aachen. 1999.

[SwKo97] Carl Swail und Robert Kobierski. Direct Voice Input for Control of an Avionics Management System. In *53rd Annual Forum of the American Helicopter Society*, Virginia Beach, 1997. American Helicopter Society.

[Tull00] Mike Tull. *Boeing JSF to Feature Voice-Recognition Technology.* Boeing News Release February 22nd. 2000.

[Vase96] Saeed V. Vaseghi. *Advanced Signal Processing and Digital Noise Reduction.* Wiley, Teubner, Stuttgart, Chichester. 1996.

[Weib01] Roland Weibel. *Human Factors and Automation Concerns for Integrating New Technology into the Airplane Flight Deck.* Department of Aerospace Engineering, University of Kansas. 2001.

[West01] Martin Westphal. *Robuste kontinuierliche Spracherkennung für mobile Informationssysteme.* Shaker Verlag, Aachen. Dissertation an der Universität Karlsruhe (TH), 2001.

[WiBL96] David T. Williamson, Tomothy P. Barry und Kristen K. Liggett. Flight Test Results Of ITT VRS-1290 In NASA OV-10. In *Proceedings of 15th Annual International Voice Technologies Applications Conference*, San Jose, 1996. American Voice Input/Output Society, S. 33–40.

[Will97] David. T. Williamson. Robust Speech Recognition Interface to the Electronic Crewmember: Progress and Challenges. In *Proceedings of 4th Human-Electronic Crewmember Workshop*, Kreuth, Germany, 1997.

[WiSt85] Bernard Widrow und Samuel D. Stearns. *Adaptive Signal Processing.* Prentice Hall, Englewood Cliffs, New Jersey. 1985.

[Wölf03] Matthias Wölfel. *Minimum Variance Distortionless Response Spectral Estimation and Subtraction for Robus Speech Recognition.* Diplomarbeit an der Universität Karlsruhe (TH) und Carnegie Mellon University Pittsburgh. 2003.