ROBUSTE SPRACHERKENNUNG IM COCKPIT VON LUFTFAHRZEUGEN

Michael Dambier, Matthias Wölfel und Christian Fügen

Interactive Systems Laboratories
Institut für Logik, Komplexität und Deduktionssysteme
Universität Karlsruhe (TH)

{*mdambier, wolfel, fuegen*}*@ira.uka.de*

Kurzfassung: Der Einsatz von Spracherkennung im Luftfahrzeug kann dazu beitragen, die Arbeitsbelastung des Piloten zu reduzieren. In Flugphasen mit hoher Arbeitsbelastung können komplexe Abläufe mit Hilfe der Spracheingabe vereinfacht werden. Denkbar wären eine Sprachsteuerung von Kommunikations- und Navigationseinrichtungen; weiterhin eine automatische Transkription und Darstellung von Funksprüchen der Flugverkehrskontrollstellen. Um einen zusätzlichen Trainingsaufwand der Piloten eines Luftfahrzeuges zu verhindern, sollte eine kontinuierliche, sprecherunabhängige Spracherkennung eingesetzt werden. Die Verwendung von Standardverfahren ist in diesem Umfeld jedoch nicht geeignet, da die Sprache durch im Cockpit vorhandene Störgeräusche beeinflusst wird. Abhängig von Antrieb und Flugphase des Luftfahrzeuges können diese Störgeräusche sehr unterschiedlich sein. Ein weiteres Problem stellt die im Flugfunk verwendete und sehr störanfällige Amplitudenmodulation dar. In der hier vorgestellten Arbeit wird untersucht, inwieweit Verfahren, die zur Spracherkennung in Kraftfahrzeugen entwickelt wurden, für die Spracherkennung in Luftfahrzeugen der Allgemeinen Luftfahrt geeignet sind. Betrachtet werden die spektrale Subtraktion, cepstrale Mittelwertsubtraktion und die modellkombinationsbasierte akustische Transformation. Des weiteren werden diese Verfahren adaptiert, um sie auf die Besonderheiten im Cockpit von Luftfahrzeugen anzupassen. Sprachaufnahmen, aufgenommen in einem Kleinflugzeug und einem Helikopter, wurden mit dem Spracherkenner *Janus Recognition Toolkit* (JRTk) der *Interactive Systems Labs* (Universität Karlsruhe (TH), Deutschland und Carnegie Mellon University, Pittsburgh, USA) automatisch transkribiert. Im Kleinflugzeug wurde eine Reduktion der Wortfehlerrate von 25,47% auf 19,30%, im Helikopter von 28,86% auf 17,45% erreicht.

1 Einleitung

Spracherkennung im Cockpit von Luftfahrzeugen kann helfen, die Arbeitsbelastung des Piloten zu verringern und Flugunfälle zu vermeiden. Leider sind die Kosten der Aufnahme mehrerer Stunden Trainingsmaterial für ein spezielles Cockpit-Spracherkennungssystem sehr hoch, so dass auf Akustiken reiner Sprache (”Laborakustik”) zurückgegriffen werden muss. In dieser Arbeit wurden deshalb signalbasierte Verfahren zur Geräuschreduktion auf Ihren Einsatz in einem Cockpit-Spracherkennungssystem mit ”Laborakustik” überprüft.

Forschung an Spracherkennungssystemen 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 im allgemeinen eine höhere Erkennungsrate 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 Satzerkennungsraten von circa 97 Prozent bei einem relativ kleinen Vokabular (circa 50 - 100 Wörter) [3][4][5]. Dies ist durch den Kommandocharakter der Spracheingaben bedingt, wobei die ”Eingabesätze” meist weniger als drei Worte umfassen.

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, des Flight Management Systems (FMS), 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 damit erhöht werden. Durch eine korrekte Funktion können auch Einstellungsfehler der Flugzeugbesatzung, welche unbemerkt zu einem Flugunfall führen können, vermieden werden.

Da in der Allgemeinen Luftfahrt Piloten häufig das Flugzeug wechseln, ist eine sprecherabhängige Spracherkennung und die Verwendung von vorgegebenen Kommandosätzen nicht praktikabel. Deshalb untersuchten wir in dieser Arbeit den Einsatz verschiedener Vorverarbeitungsverfahren aus der Spracherkennung in Kraftfahrzeugen in einem sprecherunabhängigen, zur Erkennung spontaner Sprache konzipierten Spracherkennungssystem im Cockpit eines Luftfahrzeuges. Aus der großen Anzahl möglicher Verfahren wählten wir solche Verfahren aus, die bereits bei der Spracherkennung in Kraftfahrzeugen Verbesserungen erzielt haben. Diese wurden durch die Verwendung einer neuen Sprach-Pause-Detektion für einen Einsatz im Cockpit verbessert. Zur Verbesserung der Spracherkennungsraten sollten keine künstlich verrauschten Daten zum Training eines neuen Spracherkenners benutzt werden, da dieses Vorgehen bereits erfolgreich angewendet wurde.

2 Geräuschanalyse

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 zur Erhöhung der Geräuschkulisse bei. Weiterhin variieren die Geräusche in Abhängigkeit vom Flugzustand des Luftfahrzeuges. 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. Abbildung 1 zeigt die Störgeräuschspektren sowie deren Varianz des Helikopters und des Kleinflugzeuges.

Das Störgeräuschspektrum des Helikopters zeigt zwei Spitzen im Bereich von 0 bis 500 Hz und von 500 bis 1000 Hz. Die größte Varianz des Störgeräuschs ist zwischen 0 und 400 Hz lokalisiert. Das SNR liegt bei 10,4 dB, die Varianz bei 10,6.

Das durchschnittliche Störgeräusch des Kleinflugzeugs 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 als Ursache.

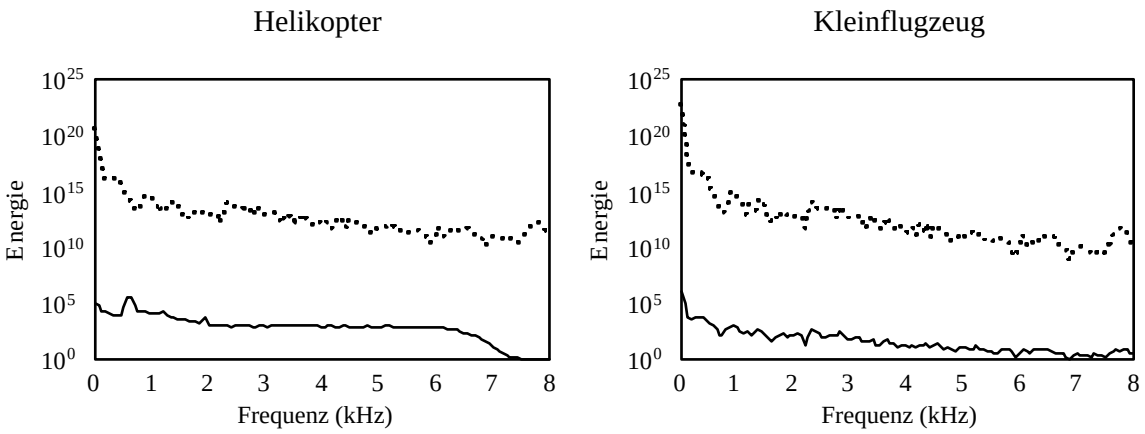

Abbildung 1 - Analyse der Störgeräusche der Luftfahrzeuge (Abbildungen mit logarithmischer Y-Achse, Störgeräuschspektrum durchgezogene Linie, Varianz gestrichelte Linie)

3 Betrachtete Verfahren

Die spektrale Subtraktion, die modellkombinationsbasierte akustische Transformation, die sprachbasierte cepstrale Mittelwertsubtraktion und eine für alle Verfahren notwendige Sprach-Pause-Detektion wurden betrachtet.

Die spektrale Subtraktion ist ein bekanntes und häufig genutztes Verfahren zur Minderung der Einflüße additiver Störgeräusche. Voraussetzung bei diesem Verfahren ist, dass das Störgeräusch das Nutzsignal nur additiv überlagert und nicht mit diesem korreliert ist. Die

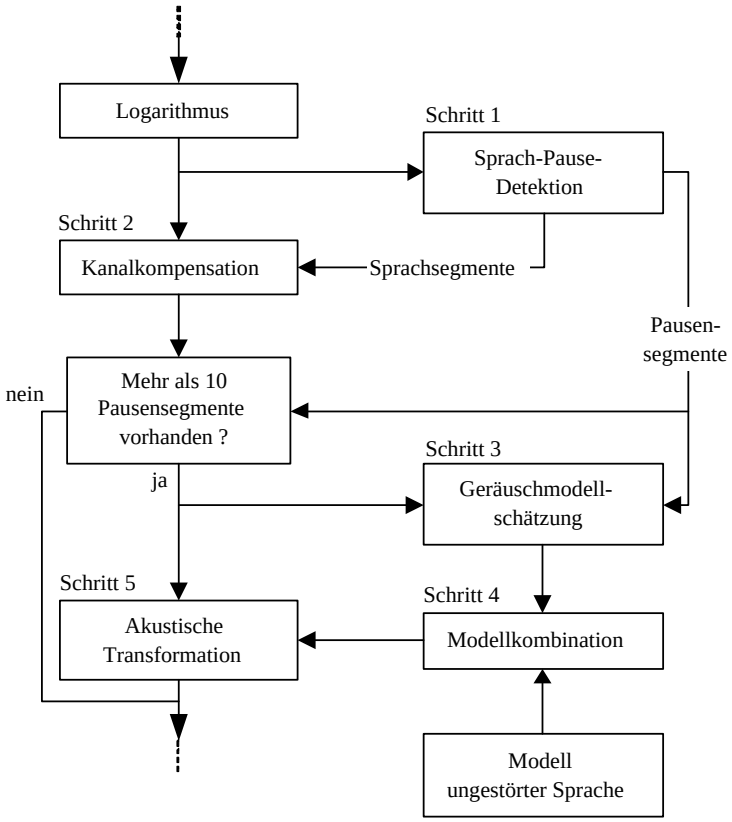

Abbildung 2 - Schema der Modellkombinationsbasierten Akustischen Transformation

Schätzung des zu subtrahierenden Störgeräusches erfolgte mit dem Verfahren der Minimum Statistik [2] auf der jeweiligen zu verarbeitenden Äußerung.

Die modellkombinationsbasierte akustische Transformation (MAM, engl.: Model Combination Based Acoustic Mapping) ist ein Verfahren zur Kompensation der Umgebungseinflüsse [9]. 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 ein konstanter Kanal für die Dauer der Aufnahme. Wichtigste Voraussetzung ist die Erstellung eines Modells ungestörter Sprache mit einem Codebuch mit 100 Gaussians auf den Trainingsdaten des Spracherkennungssystems.

Die einzelnen Schritte der MAM sind in Abbildung 2 abgebildet. Im ersten Schritt werden die Sprachpausensegmente der vorliegenden Äußerung bestimmt. Aus diesen wird dann im dritten Schritt ein Störgeräuschmodell mit einem Codebuch mit einer Gaussian erstellt. Auf den detektierten Sprachsegmenten wird im zweiten Schritt eine sprachbasierte Mittelwertsubtraktion durchgeführt. Im Schritt vier wird ein Modell gestörter Sprache aus der Kombination (PMC, engl.: Parallel Model Combination) des Störgeräuschmodells mit dem Modell ungestörter Sprache erstellt. Aus den Mittelwerten der Modelle ungestörter und gestörter Sprache wird dann im Schritt fünf eine Transformation berechnet. Diese akustische Transformation wird danach auf die Merkmalsvektoren der jeweiligen Äußerung angewandt.

Die sprachbasierte cepstrale Mittelwertsubtraktion (SCMS) ist heutzutage ein Standardver-

fahren zur Kanalkompensation und wird deshalb nicht näher erläutert. Die SCMS wurde bei allen Experimenten verwendet. Weiterhin wurden die cepstrale Mittelwertsubtraktion (CMS) sowie die 2-Level cepstrale Mittelwertsubtraktion (2CMS) betrachtet. Bei der 2CMS wird nicht nur der Mittelwert aller Sprachsegmente von diesen subtrahiert sondern auch ein Mittelwert über alle Pausensegmente gebildet und von diesen subtrahiert [9]. Mit diesen Verfahren wurde jedoch keine Verbesserung der Wortfehlerraten erzielt.

Alle bisher vorgestellten Verfahren benötigen eine Sprach-Pause-Detektion. Diese Detektion wurde im cepstralen Bereich durchgeführt, da diese gegenüber einer Detektion im Spektralbereich eine hohe Unabhängigkeit von Amplitude und Störgeräuschlevel [1] sowie das Erreichen guter Ergebnisse bei niedrigen SNR-Werten und nicht-stationären Störsignalen [7] bietet. In [1] und [7] werden verschiedene Distanzmaße für die Sprach-Pause-Detektion im Cepstralbereich vorgeschlagen. Diese wurden jedoch nicht verwendet, da das im folgenden vorgestellte und im Rahmen einer Diplomarbeit [10] entwickelte Verfahren in den Versuchen bessere Segmentierungen lieferte. Abbildung 3 zeigt die Implementierung der Sprach-Pause-Detektion in die Standard-Vorverarbeitung des Spracherkennungssystems. Mit einem einfachen energiebasierten Sprach-Pause-Detektor wurden zunächst im spektralen Bereich Pausestellen der jeweiligen Äußerung detektiert. Zur Detektion wurde ein adaptiver Schwellwert verwendet. Die Berechnungsvorschrift wurde in mehreren Experimenten als

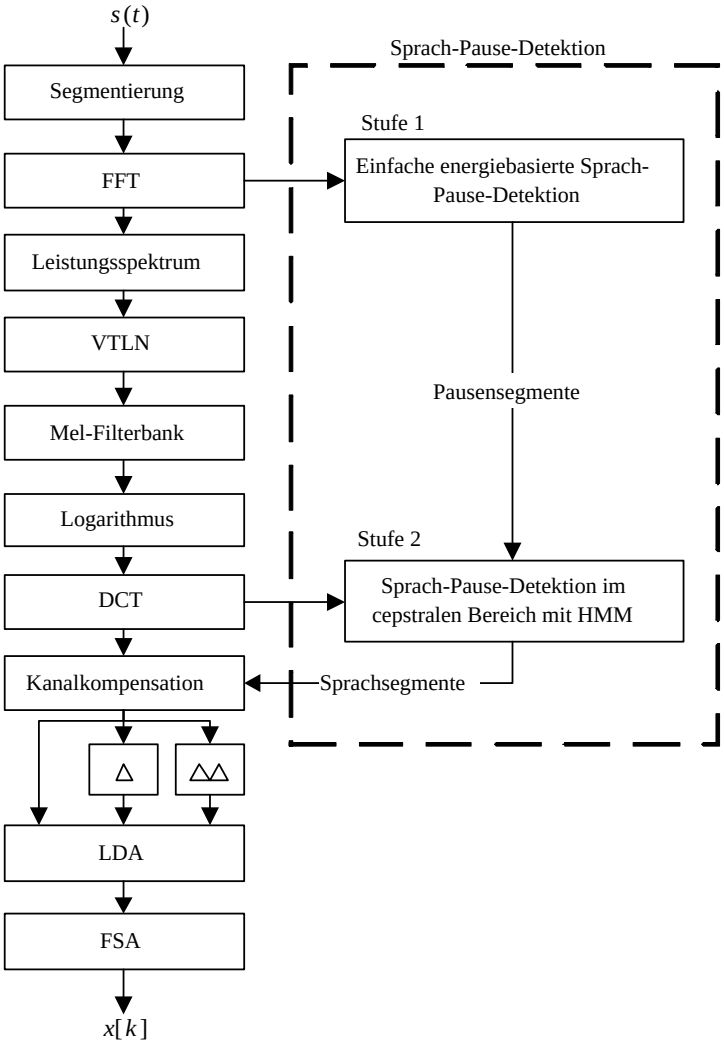

Abbildung 3 - Verbesserte Sprach-Pause-Detektion im Ablauf der Standard-Vorverarbeitung

$$v_{specThreshold} = \frac{3.0}{20 \cdot (speechMean - silMean) \cdot \log_{10} 2.0}$$

festgelegt, wobei $speechMean$ und $silMean$ die Ergebnisse des k-nächste-Nachbarn-Verfahrens, berechnet auf den Energiewerten der Analyserahmen der Äußerung, sind.

Auf diesen im Spektralbereich festgelegten Pausesegmenten der Äußerung wurde im Cepstralbereich ein Hidden Markov Model (HMM) mit einer Gaussian trainiert, da sich geringfügig ändernde Störgeräusche, wie beispielsweise Helikoptergeräusche, durch ein HMM mit einem Zustand modelliert werden können [8]. Jedem Analyserahmen der Äußerung wird ein Wahrscheinlichkeitswert $p[i]$ durch Auswertung der Gaußdichte im Cepstralbereich zugewiesen. Die Klassifikation erfolgt nach der Anwendung eines Medianfilters auf die Wahrscheinlichkeitswerte der gesamten Äußerung durch einen Vergleich mit einem weiteren adaptiven Schwellwert $v_{cepstThreshold}$. Dieser wird über alle Wahrscheinlichkeitswerte der N Analyserahmen der Äußerung berechnet.

$$v_{cepstThreshold} = \lambda \cdot \frac{1}{N} \cdot \sum_{i=0}^{N-1} p[i]$$

wobei der Skalierungsfaktor λ in mehreren Experimenten auf $\lambda = 0.8$ festgelegt wurde.

4 Experimente

Die unten aufgeführten Experimente wurden mit dem *Janus Recognition Toolkit* (JRTk) der *Interactive Systems Labs* (Universität Karlsruhe (TH), Deutschland und Carnegie Mellon University, Pittsburgh, USA) durchgeführt. Der im JRTk verwendete One-Pass-Decoder IBIS [6] ermöglicht die Verwendung einer kontextfreien Grammatik als Sprachmodell. Die verwendete Grammatik wurde bei den Versuchen von Hand erstellt. Das verwendete Vokabular umfasst 405 Wörter. Das Spracherkennungssystem besitzt 2143 Codebücher mit jeweils 16 Gaussians.

Als Trainingsmaterial für die ”Laborakustik” wurde der *Broadcast News (BN) Corpus* und der *English-Spontaneous-Scheduling-Task (ESST) Corpus* verwendet. Die Broadcast-News-Daten bestehen aus Äußerungen von Nachrichtensprechern, auch bei Aussenaufnahmen, mit entsprechender Geräuschkulisse. Die Trainingsdaten umfassen ungefähr 106 Stunden Sprache, gesprochen von 4019 Sprechern beider Geschlechter. Die Daten des English-Spontaneous-Scheduling-Task 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. Die Corpora beinhalten spontan sprachliche Äußerungen wie beispielsweise ”äh’s”. Die Daten wurden mit einer Abtastfrequenz von 16 kHz und einer Quantisierung von 16 bit aufgenommen.

4.1 Sprachdaten

Für Experimente standen Sprachdaten aus einem Helikopter vom Typ MD 520 N sowie einem Kleinflugzeug vom Typ Socata TB 20 Trinidad GT zur Verfügung. Diese wurden im

Rahmen dieser Arbeit mit Standard-Luftfahrtheadsets der Firmen Sennheiser (HME 100) und Bose (AH-TC) aufgenommen. Einen Überblick über die verwendeten Daten gibt Tabelle 1. Insgesamt standen 1299 Äußerungen zur Verfügung. Davon wurden 100 Äußerungen von zwei männlichen Sprechern im Helikopter und 1199 Äußerungen von einem männlichen Sprecher im Kleinflugzeug aufgenommen. Diese Daten wurden in Entwicklungs- und Testdaten aufgeteilt. Als Testdaten standen insgesamt 329 Äußerungen zur Verfügung. Davon wurden 30 Äußerungen im Helikopter und 299 Äußerungen im Kleinflugzeug aufgenommen.

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

Tabelle 1 - Aufteilung der Sprachdaten

4.2 Ergebnisse

In Tabelle 2 sind die Ergebnisse der Experimente aufgeführt. Dabei zeigte sich, dass die MAM der spektralen Subtraktion überlegen ist. Die Zunahme der Worfehlerrate bei der spektralen Subtraktion, angewandt auf die Sprachdaten aus dem Kleinflugzeug, läßt sich auf die sehr unterschiedlichen Störgeräusche in unterschiedlichen Flugphasen und auf die Festlegung konstanter Parameter bei der Berechnung der spektralen Subtraktion zurückführen. Ebenso die unterschiedlichen Ergebnisse der MAM auf den Daten des Kleinflugzeuges und des Helikopters.

Verfahren	Helikopter	Kleinflugzeug
Standard-Vorverarbeitung	28,86%	25,47%
Spektrale Subtraktion	24,83%	27,64%
MAM	17,45%	19,30%

Tabelle 2 - Ergebnisse der Versuche auf den Testdaten (Angaben der Wortfehlerraten in %)

Das Spracherkennungssystem mit BN/ESST-Laborakustik und einer kontextfreien Grammatik als Sprachmodell bei einer Vokabulargröße von 405 Wörtern ist für den Einsatz in beiden betrachteten Luftfahrzeugen am besten geeignet. Das Gesamtsystem inklusive der Vorverarbeitung und Sprach-Pause-Detektion dekodiert Sprachdaten auf einem Computer mit Pentium-IV-Prozessor mit 2,66 GHz Taktfrequenz mit einem Echtzeitfaktor von 0,6.

5 Zusammenfassung und Ausblick

In dieser Arbeit wurde der Einsatz eines Spracherkennungssystems mit einer sogenannten ”Laborakustik” in einem Helikopter und einem Kleinflugzeug untersucht. Durch die Verwendung der modellkombinationsbasierten akustischen Transformation (MAM) und der in dieser Arbeit vorgestellten Sprach-Pause-Detektion konnte eine durchschnittliche relative Fehlerreduktion von 31,88% erreicht werden. Die Verwendung der spektralen Subtraktion erwies sich als weniger geeignet. Dies lag vor allem daran, dass keine Berechnungsvorschrift zur adaptiven Anpassung des Spectral Floor und des Overestimation Factors der spektralen

Subtraktion verwendet wurde, sondern diese Parameter nach einigen Versuchen auf den Entwicklungsdaten als konstant festgelegt wurden.

Für weitere Arbeiten auf diesem Gebiet ist eine umfangreichere Datensammlung und die Betrachtung weiterer Verfahren zur Geräuschreduktion unerlässlich. Weitere Verbesserungen könnten durch eine flugphasen- und luftfahrzeugabhängige Störgeräuschunterdrückung und eine Adaption auf den Sprecher unter der Annahme, dass ein Luftfahrzeug der Allgemeinen Luftfahrt für die Dauer des Fluges von nur einem Piloten geflogen wird, erzielt werden. Der verfolgte Ansatz könnte auch auf den Flugfunk und somit auf Spracherkennungssysteme in der Flugsicherung übertragen werden.

Literatur

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

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

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

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

[10] Dambier, M. *Signalbasierte Verfahren zur robusten Spracherkennung im Cockpit von Luftfahrzeugen.* Diplomarbeit, Universität Karlsruhe (TH), Karlsruhe, Deutschland, November 2003.