Hochschule Niederrhein
University of Applied Sciences

# Ergebnisse des 3. Projektabschnitts

Das Ziel des dritten Projektabschnitts ist die Entwicklung eines Verfahrens zur Extraktion robuster akustischer Merkmale aus gestörten Sprachsignalen. Als Basis wird dazu ein am Institut für Kommunikationsakustik an der Ruhr-Universität Bochum entwickeltes Verfahren zur Störunterdrückung verwendet [1],[2]. Neben der Unterdrückung stationärer Störgeräusche soll das Merkmalsextraktionsverfahren die Kompensation unbekannter Frequenzgänge, die beispielsweise auf den Einsatz unterschiedlicher Mikrofone zurückgeführt werden können, gewährleisten. Zudem werden die Möglichkeiten untersucht, eine robuste Erkennung gestörter Signale, die im Freisprechmodus in Räumen aufgenommen wurden, zu realisieren. Dies wird letztlich durch die Kombination der robusten Merkmalsextraktion und einer Adaption der Referenzmuster auf die durch den Nachhall geprägte akustische Umgebung erreicht. Eine ausführliche Untersuchung der derzeit vorhandenen Ansätze einer Signalverarbeitung zur Reduktion des Halls, die zudem mit vertretbarem Aufwand in die robuste Merkmalsextraktion integriert werden kann, kam zu dem Schluss, dass es derzeit keinen derartigen Ansatz gibt, der eine Verbesserung der Erkennungsraten von im Freisprechmodus aufgenommenen Signalen ermöglicht. Die Untersuchungen und die erzielten Ergebnisse werden in einem separaten Projektbericht [3] dokumentiert.

## *Extraktion robuster akustischer Merkmale*

Das als Basis herangezogene Verfahren zur Störunterdrückung [1] beruht auf einer DFT basierten Kurzzeit-Spektralanalyse des gestörten Sprachsignals. Mit Hilfe einer Schätzung des Spektrums der als stationär angenommenen Hintergrundstörung erfolgt eine adaptive Filterung im Frequenzbereich. Die gefilterten Spektren werden in den Zeitbereich zurück transformiert zur Generierung eines störreduzierten Sprachsignals. Der neue Aspekt des betrachteten Verfahrens ist der Einsatz einer „cepstralen Glättung“ der Charakteristik des adaptiven Filters. Die Entwickler dieses Verfahrens konnten auch die Möglichkeiten einer Verbesserung der automatischen Erkennung gestörter Sprachsignale aufzeigen. Dazu wurden aus den störbefreiten Sprachsignalen mit einem separaten Verfahren die für die Erkennung relevanten Merkmale extrahiert.

Eine Aufgabe dieses Arbeitsabschnitts bestand in der Integration der auf der „cepstralen Glättung“ beruhenden Filterung in das vorhandene, im vorhergehenden Arbeitsbericht bereits dargestellte Verfahren „HGH“ zur Extraktion der Energie und der cepstralen Merkmale. Die

Rücktransformation in den Zeitbereich, wie sie auch in dem von ETSI standardisierten Verfahren zur Extraktion robuster Merkmale stattfindet, soll dabei vermieden werden. Zudem werden im Vergleich zu den Störunterdrückungsverfahren bei der Sprachanalyse zur Erkennung kürzere Signalabschnitte bei einer ebenfalls geringeren Verschiebung aufeinanderfolgender Analysefenster betrachtet.

Eine Gegenüberstellung des „HGH“ Extraktionsverfahrens und des neuen Verfahrens zur Extraktion robuster Merkmale wird in Bild 1 vorgenommen.

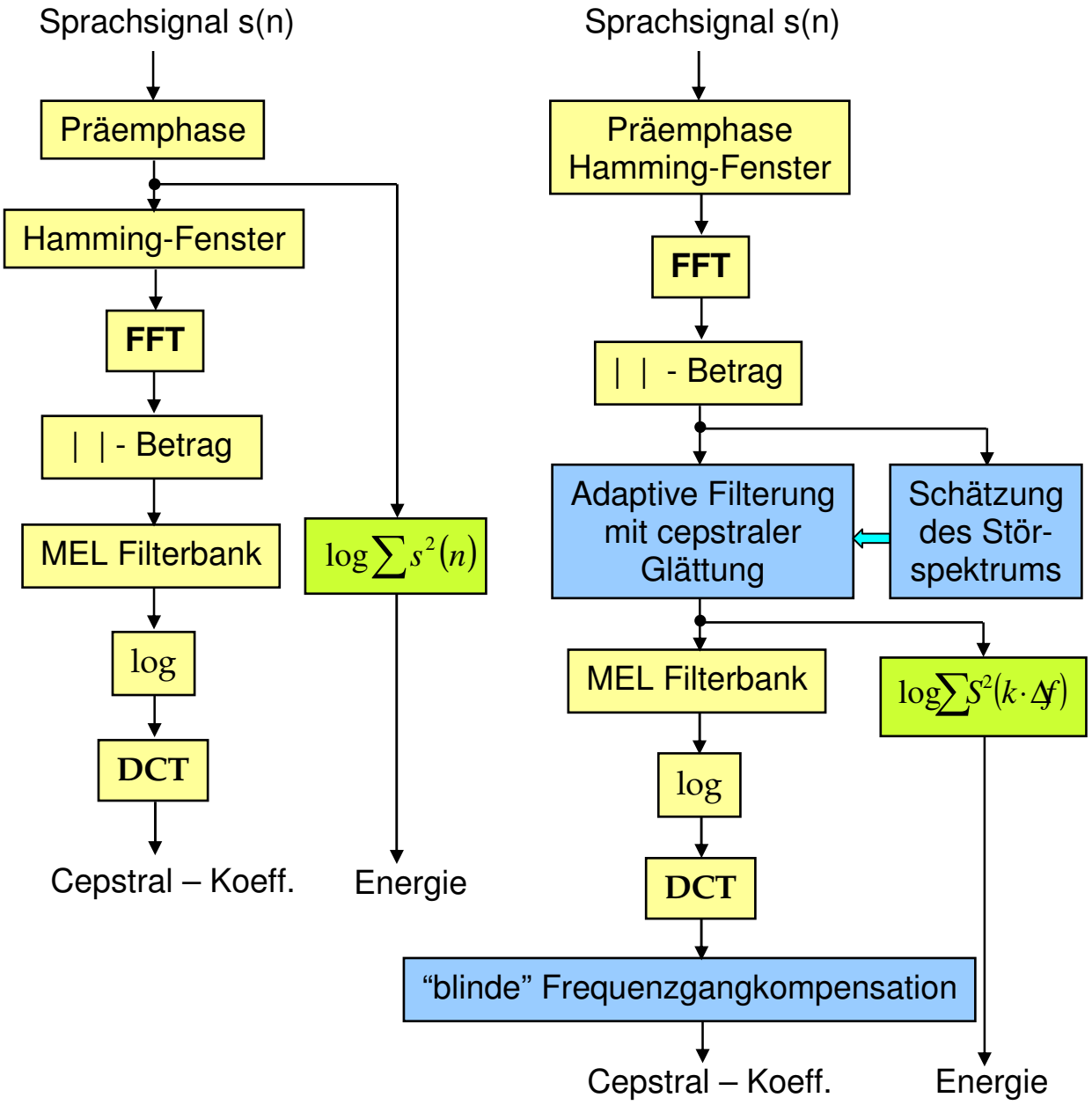


Bild 1: Gegenüberstellung der Merkmalsextraktionsverfahren

Bis zur Betragsbildung des DFT Spektrums sind beide Verfahren gleich. Es werden jeweils Signalabschnitte mit einer Länge von 25 ms, entsprechend 200 Abtastwerten bei einer Abtastfrequenz von 8 kHz, und einer Verschiebung des Analysefensters um 10 ms transformiert. Die Schätzung des Störspektrums wird mit Hilfe eines aus einem früheren Ansatz [8] weiterentwickelten Verfahrens vorgenommen. Eine genaue mathematische Beschreibung der adaptiven Filterung mit cepstraler Glättung findet sich im Anhang. Im Gegensatz zur HGH Analyse, bei der die Energie eines Signalabschnitts aus den quadrierten Abtastwerten des höhenangehobenen Zeitsignals bestimmt wird, wird der Energieparameter aus den quadrierten Werten des gefilterten Spektrums berechnet. Die Bestimmung des logarithmierten Mel-Spektrums und der Cepstralparameter ist bei beiden Verfahren identisch. Zur „blinden“ Kompensation unbekannter Frequenzgänge wird abschließend das durch die Cepstral-Koeffizienten beschriebene Spektrum mit einem „mittleren“ Sprachspektrum bzw. den zugehörigen Cepstral-Koeffizienten verglichen [4]. Diese Kompensation entspricht der Vorgehensweise in dem von ETSI standardisierten Verfahren [5]. Eine mathematische Beschreibung findet sich im Anhang.
Der logarithmierte Energiewert und die Cepstral-Koeffizienten jedes analysierten Signalabschnitts werden um die Delta und Delta-Delta Parameter ergänzt, die gemäß der Vorgehensweise, wie sie im von ETSI standardisierten Verfahren [5] festgelegt wurde, bestimmt werden.

### *Erkennungsexperimente mit der „Aurora-2“ Datensammlung*

Zur Verifikation der mit dem robusten Merkmalsextraktionsverfahren erzielten Ergebnisse wurden zunächst Erkennungsexperimente mit der unter der Bezeichnung „Aurora-2“ [6] bekannten Sammlung gestörter Sprachsignale durchgeführt, da die Autoren von [2] Erkennungsraten für dieses Experiment angeben. Wie bei der im vorhergehenden Bericht erwähnten „Aurora-5“ Sammlung wurden die unter der Bezeichnung TIDigits bekannten Aufnahmen englischer Ziffern als Basis zur Erzeugung der gestörten Aurora-2 Daten verwendet. Im Vergleich zu Aurora-5 wird bei jeder Testbedingung eine deutlich geringere Anzahl von Aufnahmen benutzt. Zudem wurden die überlagerten Störsignalabschnitte teilweise nur recht kurzen Aufnahmen der jeweiligen Hintergrundstörung entnommen. Des Weiteren beinhaltet Aurora-2 keine Aufnahmen im Freisprechmodus.
Die Fehlerraten, die sich bei Einsatz der robusten Merkmalsextraktion im Vergleich zu den bei [1] und bei Verwendung des robusten ETSI-2 Verfahrens erzielten Ergebnissen einstellen, können der nachstehenden Tabelle entnommen werden. Dabei wurden zunächst nur die als „Set A“

bezeichneten Störsignale betrachtet. Die angegebenen Prozentwerte repräsentieren mittlere Fehlerraten für den SNR Bereich von 0 bis 20 dB. Dabei wird für die SNR Werte von 0, 5, 10, 15 und 20 dB jeweils die Erkennung von 1001 Äußerungen ausgewertet.

| Verfahren | Störung | | | | Mittelwert |
|---|---|---|---|---|---|
| | „subway“ | „babble“ | „car“ | „exhibition“ | |
| Experimente in [1] | 13,61 % | 15,90 % | 12,50 % | 14,33 % | 14,08 % |
| Robuste Merkmale (ohne Frequenzgangkompensation) | 15,39 % | 17,65 % | 12,41 % | 13,81 % | 14,82 % |
| Robuste Merkmale (mit Frequenzgangkompensation) | 16,13 % | 17,41 % | 11,78 % | 15,02 % | 15,09 % |
| ETSI-2 | 11,72 % | 15,97 % | 9,22 % | 12,11 % | 12,25 % |

Tabelle 1: Mittlere Wortfehlerraten für Set-A des „Aurora-2“ Experiments

Die mittlere Fehlerrate über alle Störbedingungen ist bei der robusten Merkmalsextraktion mit einer Kompensation eines unbekannten Frequenzgangs um etwa 1 % schlechter im Vergleich zu den bei [1] angegebenen Werten und um knapp 3 % schlechter im Vergleich zur Merkmalsextraktion gemäß des ETSI Standards. Bei einer genaueren Analyse der Erkennungsergebnisse stellt man fest, dass die Verschlechterung im Vergleich zu den mit dem ETSI-2 Verfahren erzielten Ergebnissen im Wesentlichen auf schlechtere Ergebnisse bei einem SNR von 0 dB zurückzuführen ist. Bei höheren SNR Werten stellt sich meist sogar eine Verbesserung ein.

Der Einsatz der blinden Frequenzgangkompensation führt bei den Experimenten mit den Sprachdaten des Sets-A zu einer geringfügigen Verschlechterung der Erkennungsergebnisse, wobei die Signale des Sets-A nur durch die additive Überlagerung von Störgeräuschen generiert wurden und der Einfluss eines unbekannten Frequenzgangs nicht berücksichtigt wurde. In der nachstehenden Tabelle sind auch die mittleren Ergebnisse für „Set-B“ und „Set-C“ enthalten. Set-B ist durch die Verwendung anderer Störgeräusche und Set-C durch die zusätzliche Berücksichtigung einer Frequenzcharakteristik, mit der die Übertragung über einen Telefonkanal simuliert wird, gekennzeichnet.

| Verfahren | Set-A | Set-B | Set-C |
|---|---|---|---|
| Robuste Merkmale (ohne Frequenzgangkompensation) | 14,82 % | 15,89 % | 20,70 % |
| Robuste Merkmale (mit Frequenzgangkompensation) | 15,09 % | 15,34 % | 18,81 % |
| ETSI-2 | 12,25 % | 12,90 % | 13,97 % |

Tabelle 2: Mittlere Wortfehlerraten für alle Sets des „Aurora-2“ Experiments

Die Erkennungsergebnisse mit der auf der cepstralen Glättung beruhenden robusten Merkmalsextraktion sind um einige Prozentpunkte schlechter als die Fehlerraten, die mit dem von ETSI standardisierten Verfahren erzielt werden. Die Verschlechterung ist im Wesentlichen auf höhere Fehlerraten bei einem geringeren SNR zurückzuführen. Zudem ist zu berücksichtigen, dass das standardisierte Verfahren während der vergleichenden Untersuchungen zur Festlegung des Standards über Jahre hinweg auf dieses Erkennungsexperiment hin optimiert wurde. Bei den im nachfolgenden Abschnitt dargestellten Erkennungsexperimenten ergaben sich in vielen Fällen schlechtere Ergebnisse mit dem standardisierten Verfahren.

### *Ergebnisse für die im 2.Projektabschnitt festgelegten Experimente*

Im Weiteren wurden mit der auf der cepstralen Glättung beruhenden robusten Merkmalsextraktion, die nachfolgend mit dem Kürzel HGH-Robust gekennzeichnet wird, Erkennungsexperimente für die im zweiten Projektabschnitt definierten Experimente durchgeführt. Für die Sprachdaten der „Aurora-5“ Sammlung, die die Aufnahme in der gestörten Umgebung eines Autos im Freisprechmodus berücksichtigt, wurden die in Tabelle 3 enthaltenen und in Bild 2 visualisierten Fehlerraten erzielt. Es werden die Ergebnisse bei Einsatz der von ETSI standardisierten Merkmalsextraktion mit den Fehlerraten verglichen, die sich bei Anwendung der robusten Merkmalsextraktion HGH-Robust bzw. für die HGH Sprachanalyse (Cepstralanalyse mit 24 Mel Bändern) in Kombination mit einer Adaption der HMMs einstellen.

Das auf der robusten Merkmalsextraktion HGH-Robust beruhende Verfahren erzeugt bis auf den Fall des SNRs von 0 dB bessere Ergebnisse als das standardisierte Verfahren. Mit der Adaption der HMMS lassen sich bei den SNRs von 10 und 15 dB die besten Ergebnisse erzielen.

| | | SNR/dB | | | | |
|---|---|---|---|---|---|---|
| Analyseverfahren | Adaption | Clean | 15 | 10 | 5 | 0 |
| ETSI-2 | | 0,55 % | 4,17 % | 7,62 % | 15,42 % | 33,48 % |
| HGH-Robust | | 0,54 % | 3,00 % | 5,99 % | 14,57 % | 35,66 % |
| HGH | X | 0,53 % | 2,09 % | 4,93 % | 15,26 % | 41,74 % |

Tabelle 3: Wortfehlerraten für „Aurora-5“ in der akustischen Umgebung „car noise handsfree“

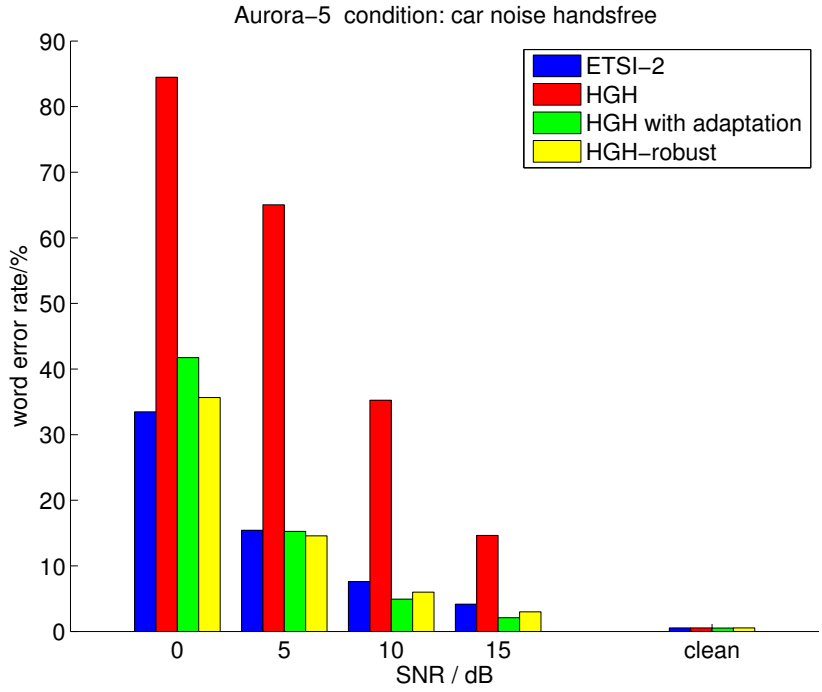


Bild 2: Wortfehlerraten für „Aurora-5“ in der akustischen Umgebung „car noise handsfree“

Ähnliche Ergebnisse stellen sich für die Erkennung der Sprachdaten ein, bei denen die Aufnahme in einer gestörten räumlichen Umgebung im Freisprechmodus simuliert wird. Die Fehlerraten können Tabelle 4 und der Darstellung in Bild 3 entnommen werden.

Die besten Ergebnisse werden bei Einsatz der HMM Adaption erzielt, da in diesem Fall der Einfluss des Nachhalls noch bis zu einem gewissen Grad kompensiert wird.

| | | SNR/dB | | | |
|---|---|---|---|---|---|
| Analyseverfahren | Adaption | Clean | 15 | 10 | 5 |
| ETSI-2 | | 0,55 % | 9,84 % | 15,98 % | 29,08 % |
| HGH-Robust | | 0,54 % | 7,83 % | 13,90 % | 27,64 % |
| HGH | X | 0,53 % | 6,38 % | 12,46 % | 27,15 % |

Tabelle 4: Wortfehlerraten für „Aurora-5“ bei Aufnahme in einer gestörten räumlichen Umgebung

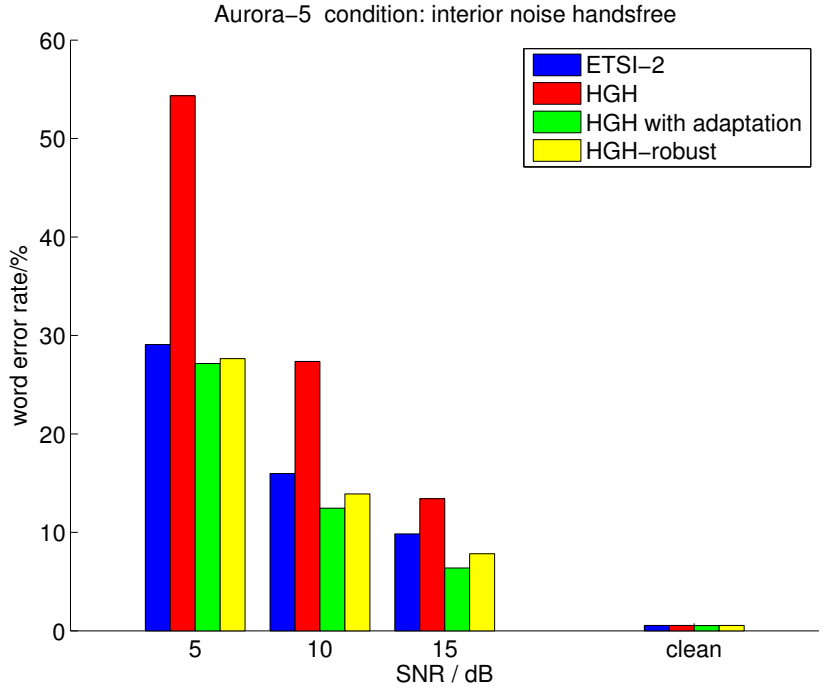


Bild 3: Wortfehlerraten für „Aurora-5“ bei Aufnahme in einer gestörten räumlichen Umgebung

Die Tabellen 5 und 6 sowie das Bild 4 enthalten die Fehlerraten für die beiden Experimente mit den RVG Daten, bei denen ebenfalls das Freisprechen im Auto und das Freisprechen in einer räumlichen Umgebung simuliert werden. Es stellen sich vergleichbare Ergebnisse ein bis auf den Fall des Freisprechens im Raum, bei dem die HMM Adaption zu einer deutlichen Verbesserung führt.

| | | SNR/dB | | | | |
|---|---|---|---|---|---|---|
| Analyseverfahren | Adaption | Clean | 15 | 10 | 5 | 0 |
| ETSI-2 | | 3,18 % | 6,02 % | 9,96 % | 16,41 % | 31,84 % |
| HGH-Robust | | 3,37 % | 6,15 % | 8,96 % | 15,94 % | 31,41 % |
| HGH | X | 3,43 % | 6,22 % | 8,97 % | 15,99 % | 30,77 % |

Tabelle 5: Wortfehlerraten für „RVG“ in der akustischen Umgebung „car noise handsfree“

| | | SNR/dB | | | |
|---|---|---|---|---|---|
| Analyseverfahren | Adaption | Clean | 15 | 10 | 5 |
| ETSI-2 | | 3,18 % | 33,56 % | 40,71 % | 52,99 % |
| HGH-Robust | | 3,37 % | 33,37 % | 40,58 % | 54,86 % |
| HGH | X | 3,43 % | 26,99 % | 37,15 % | 53,91 % |

Tabelle 6: Wortfehlerraten für „RVG“ bei Aufnahme in einer gestörten räumlichen Umgebung

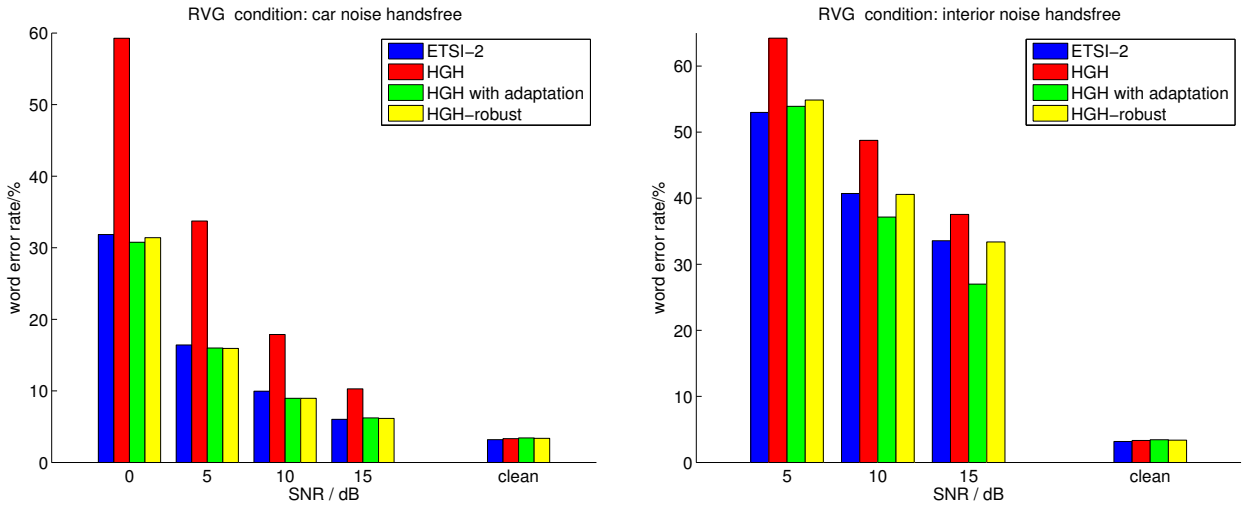


Bild 4: Wortfehlerraten für „RVG“ im Auto und in einer gestörten räumlichen Umgebung

Die Tabellen 7 und 8 enthalten die Fehlerraten für die beiden Experimente mit real in gestörten Umgebungen aufgenommenen Signalen. Bei den in Räumen aufgenommenen Sprachdaten der SpeeCon Datenbank stellen sich wiederum die besten Ergebnisse bei einer Kompensation des Nachhalls durch die Adaption der HMMs ein.

| Analyseverfahren | Adaption | Nahbesprechung | Freisprechen |
|---|---|---|---|
| ETSI-2 | | 5,48 % | 10,24 % |
| HGH-Robust | | 6,59 % | 13,12 % |
| HGH | X | 5,19 % | 13,19 % |

Tabelle 7: Wortfehlerraten für „SpeechDat-Car“

| Analyseverfahren | Adaption | Nahbesprechung | Position-1 | Position-2 | Position-3 |
|---|---|---|---|---|---|
| ETSI-2 | | 0,51 % | 5,50 % | 24,92 % | 48,65 % |
| HGH-Robust | | 0,56 % | 5,89 % | 27,67 % | 41,08 % |
| HGH | X | 1,07 % | 4,49 % | 22,22 % | 39,17 % |

Tabelle 8: Wortfehlerraten für die Erkennung der isoliert gesprochenen „SpeeCon“ Wörter

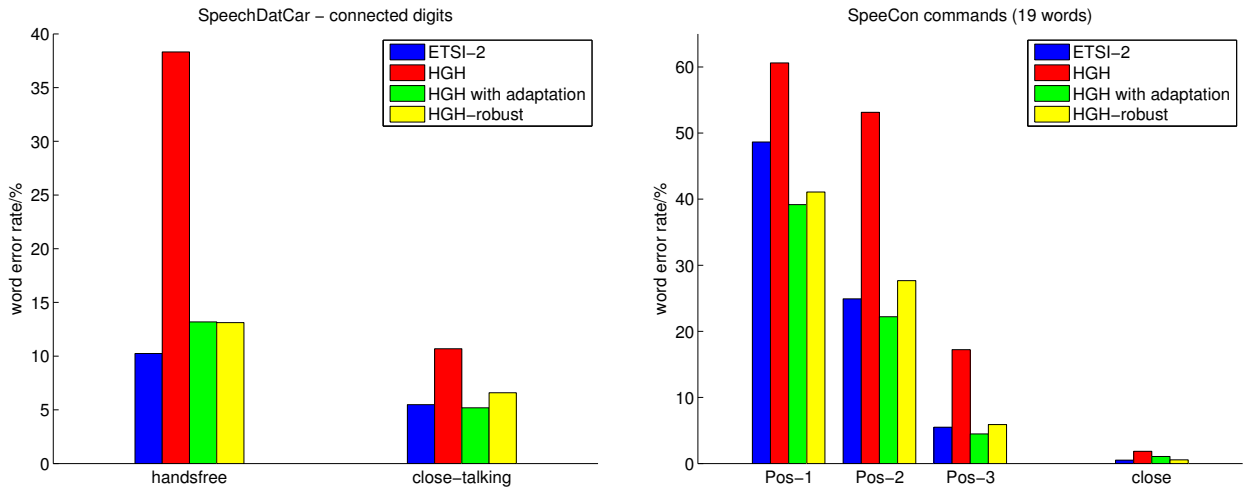


Bild 5: Wortfehlerraten für „SpeechDatCar“ und „SpeeCon“

### *Kombination mit einer Adaption der HMMs auf eine hallige, akustische Umgebung*

Im Rahmen einer ausführlichen Untersuchung der derzeit vorhandenen Ansätze einer Signalverarbeitung zur Reduktion des Halls, die zudem mit vertretbarem Aufwand in die robuste Merkmalsextraktion integriert werden kann, wurde kein verwertbarer Ansatz gefunden [3]. Daher

wurde die Möglichkeit einer Kombination der robusten Merkmalsextraktion mit einer Adaption [7] der HMMs auf den Nachhall bei der Spracheingabe im Freisprechmodus in einem Raum untersucht, wie es in Bild 6 veranschaulicht wird. Basierend auf einer Schätzung der Nachhallzeit werden dabei individuell bei jeder Spracheingabe die Mittelwerte der Cepstralparameter und der Delta und Delta-Delta Parameter, die in den HMMs enthalten sind, auf die Veränderung der zeitlichen Struktur der Sprache in Folge des Nachhalls angepasst. Die Schätzung der Nachhallzeit erfolgt dabei iterativ nach einer Spracherkennung mit Hilfe einer Maximierung der bei einer wiederholten Erkennung berechneten Wahrscheinlichkeit.

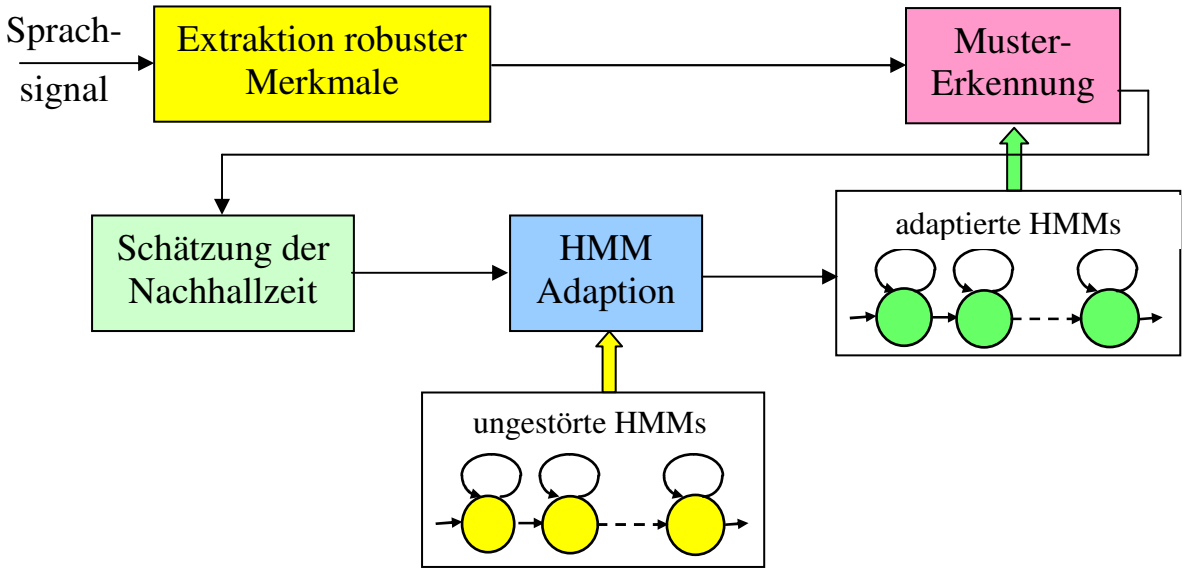


Bild 6: Kombination der robusten Merkmalsextraktion mit einer Adaption der HMMs auf Nachhall

Für die Erkennung in einer gestörten Büro-Umgebung, für die die Fehlerraten in Tabelle 4 aufgeführt wurden, stellen sich bei einer Erkennung ohne und mit Adaption die in Tabelle 9 aufgeführten Ergebnisse ein. Dabei wurde zusätzlich der Fall einer Aufnahme im Freisprechmodus („hands-free“) betrachtet, bei der keine Störung im Hintergrund auftritt. Gerade in diesem Fall macht sich der positive Effekt durch die zusätzliche Adaption der HMMs in Form einer Verringerung der Fehlerrate bemerkbar. Überlagert sich zusätzlich ein Störgeräusch, so stellt sich nur eine geringfügige Verbesserung ein. Zudem wird der Gewinn mit schlechter werdendem SNR immer geringer. Man gewinnt die Erkenntnis, dass bei einem geringen SNR die additive Überlagerung eines Störgeräuschs den gegenüber dem Nachhall dominanten Effekt darstellt.

| | | SNR/dB | | | |
|---|---|---|---|---|---|
| Analyseverfahren | Adaption | hands-free | 15 | 10 | 5 |
| HGH-Robust | | 4,36 % | 7,83 % | 13,90 % | 27,64 % |
| HGH-Robust | X | 3,30 % | 6,92 % | 13,02 % | 27,85 % |

Tabelle 9: Wortfehlerraten für „Aurora-5“ bei Aufnahme in einer gestörten Büro-Umgebung

Noch deutlicher wird der durch die Halladaption zu erzielende Gewinn bei der Betrachtung einer räumlichen Umgebung mit einer höheren Nachhallzeit. Die Fehlerraten für die Simulation der Spracheingabe in einem Wohnzimmer können Tabelle 10 entnommen werden.

| | | SNR/dB | | | |
|---|---|---|---|---|---|
| Analyseverfahren | Adaption | hands-free | 15 | 10 | 5 |
| HGH-Robust | | 9,38 % | 15,59 % | 24,23 % | 40,40 % |
| HGH-Robust | X | 5,95 % | 11,70 % | 19,55 % | 37,26 % |

Tabelle 10: Wortfehlerraten für „Aurora-5“ bei Aufnahme in einer gestörten Wohnzimmer-Umgebung

Die relative Verringerung der Fehlerraten fällt in diesem Fall höher aus, wobei man wiederum für ein schlechter werdendes SNR eine immer geringfügig werdende Verbesserung feststellt.

### *Kombination mit einer Adaption der HMMs auf eine stationäre Hintergrundstörung*

Einige frei wählbare Parameter, die bei dem zur Extraktion der robusten Merkmale eingesetzten Störunterdrückungsverfahren verwendet werden, wurden unter der Zielsetzung möglichst hoher Erkennungsraten festgelegt. Ein wichtiger Parameter ist dabei beispielsweise die Festlegung der maximalen Dämpfung des Störgeräuschs (in dB). Üblicherweise wird der Wert so gewählt, dass ein geringfügiger Anteil des Störsignals in dem störbefreiten Signal verbleibt. Dadurch wird die Hörbarkeit der bei der adaptiven Filterung auftretenden Störartefakte, die sogenannten „musical tones“ vermindert. Im Rahmen der Experimente zur Optimierung der Erkennungsergebnisse ergab sich der Wert dieser maximalen Dämpfung derart, das noch ein gewisser Anteil des Störsignals im

Spektrum des Signals nach der adaptiven Filterung erhalten bleibt. Darauf können auch die bei einigen Experimenten festgestellten höheren Fehlerraten bei einem geringen SNR zurückgeführt werden, die sich im Vergleich zu der bei ETSI standardisierten Merkmalextraktion einstellten.
Daher wurde die Möglichkeit einer Adaption der HMM Parameter auf die noch verbleibende Reststörung untersucht. Im Rahmen früherer Untersuchungen [7] wurde ein Ansatz zur Adaption von HMMs auf ein stationäres Hintergrundstörgeräusch entwickelt, für den eine Schätzung des Störspektrums benötigt wird. Da in dem Verfahren zur adaptiven Filterung ohnehin schon eine Schätzung des Störspektrums vorgenommen wird, wird dieses geschätzte Spektrum unter Berücksichtigung des festgelegten Werts für die maximale Dämpfung zur HMM Adaption verwendet.
In den Tabellen 11 und 12 werden weitere Ergebnisse für die auch im vorherigen Abschnitt (Tabellen 9 und 10) betrachteten Experimente zur Erkennung der gestörten „Aurora-5“ Daten im Freisprechmodus aufgeführt. Dabei handelt es sich um Fehlerraten, die sich bei einer ausschließlichen Adaption auf eine stationäre Hintergrundstörung ohne die im vorherigen Abschnitt angeführte Halladaption bzw. einer kombinierten Adaption auf Störung und Nachhall einstellten.
Betrachtet man zunächst die Ergebnisse bei einer alleinigen Adaption auf eine Hintergrundstörung im Vergleich zu den Ergebnissen ohne jegliche Adaption, so kann man in allen Fällen eine Verringerung der Fehlerraten feststellen. Zu schlechter werdendem SNR tritt diese Verbesserung

| | | SNR/dB | | | |
|---|---|---|---|---|---|
| Analyseverfahren | Adaption | hands-free | 15 | 10 | 5 |
| HGH-Robust | | 4,36 % | 7,83 % | 13,90 % | 27,64 % |
| HGH-Robust | auf Hall | 3,30 % | 6,92 % | 13,02 % | 27,85 % |
| HGH-Robust | auf Störung | 4,22 % | 6,70 % | 11,68 % | 22,81 % |
| HGH-Robust | auf Störung und Hall | 3,34 % | 6,82 % | 12,48 % | 25,04 % |

Tabelle 11: Wortfehlerraten für „Aurora-5“ bei Aufnahme in einer gestörten Büro-Umgebung

| | | SNR/dB | | | |
|---|---|---|---|---|---|
| Analyseverfahren | Adaption | hands-free | 15 | 10 | 5 |
| HGH-Robust | | 9,38 % | 15,59 % | 24,23 % | 40,40 % |
| HGH-Robust | auf Hall | 5,95 % | 11,70 % | 19,55 % | 37,26 % |
| HGH-Robust | auf Störung | 8,95 % | 12,22 % | 19,06 % | 32,96 % |
| HGH-Robust | auf Störung und Hall | 5,98 % | 11,44 % | 18,78 % | 34,00 % |

Tabelle 12: Wortfehlerraten für „Aurora-5“ bei Aufnahme in einer gestörten Wohnzimmer-Umgebung

immer deutlicher hervor. Vergleichbare Ergebnisse stellen sich auch für Aufnahmesituationen ein, in denen nur eine Hintergrundstörung vorhanden ist, ohne die Aufnahme im Freisprechmodus zu betrachten. Dazu finden sich weitere Ergebnisse im Anhang.
Die positive Wirkung einer Kombination der Adaptation auf Störung und Hall wird insbesondere in den Fällen keines oder eines nur geringfügigen Störgeräuschs im Hintergrund deutlich. Bei einem geringen SNR scheint der Einfluss der additiven Störung gegenüber dem Einfluss des Nachhalls zu dominieren. Es erscheint in diesen Fällen sinnvoll, nur eine Adaption auf die Hintergrundstörung, aber keine Adaption auf Hall vorzunehmen. Dies ließe sich einfach durch ein Abschalten der Halladaption bei Unterschreiten eines bestimmten SNRs realisieren.

Abschließend werden die in der gestörten Büro Umgebung erzielten Ergebnisse, die in Tabelle 11 zusammengestellt wurden, noch einmal verglichen, mit den Fehlerraten, die bei Einsatz der robusten ETSI Merkmalsextraktion und die bei Verwendung der HGH Cepstralanalyse in Kombination mit einer HMM Adaption erzielt wurden. Bild 7 beinhaltet den Vergleich der verschiedenen Verfahren. Grundsätzlich lässt sich feststellen, dass mit dem standardisierten Verfahren in allen Fällen die schlechtesten Ergebnisse erzielt werden, da dieses Verfahren nicht für die Erkennung von im Freisprechmodus aufgenommenen Daten ausgelegt ist. Speziell bei einem geringen SNR lässt sich durch die Kombination einer robusten Merkmalsextraktion und

einer Adaption auf die verbleibende Reststörung noch eine deutliche Verbesserung gegenüber allen anderen Ansätzen erzielen.

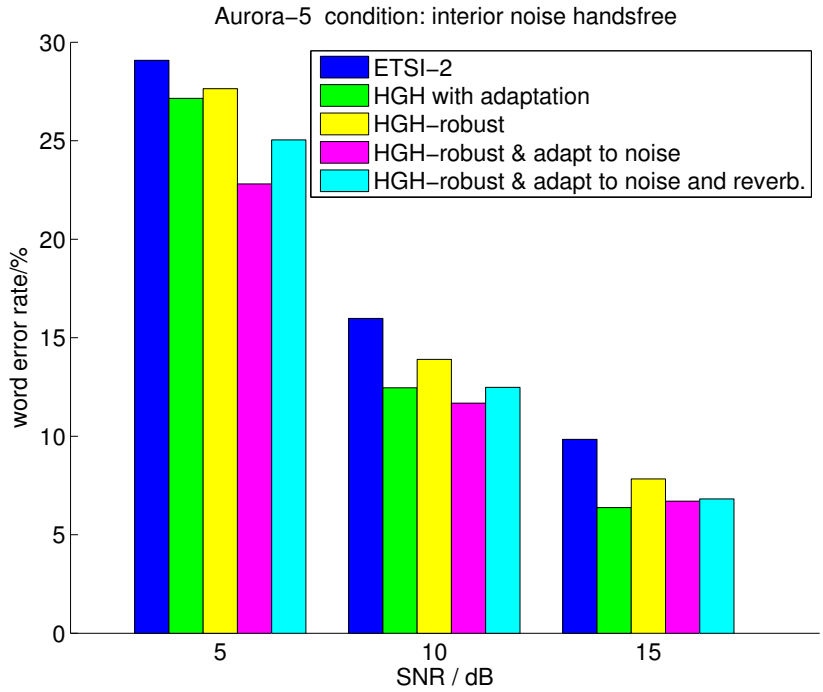


Bild 7: Wortfehlerraten für „Aurora-5“ bei Aufnahme in einer gestörten Büro Umgebung

### *Literatur*

[1] C. Breithaupt: ”Noise reduction algorithms for speech communications – Statistical analysis and improved estimation procedures”, Dissertation an der Ruhr-Universität Bochum, 2008

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

[3] A. Kitzig: “Filterung der Kurzzeit-Energieverläufe in Teilbändern zur Verbesserung der Spracherkennung bei Freisprechen“, Projektbericht, verfügbar unter http://dnt.kr.hsnr.de, 2009

[4] L. Mauuary: “Blind Equalization in the Cepstral Domain for Robust Telephone Based Speech Recognition”, European Signal Processing Conference, S. 359-362, 1998

[5] ETSI standard document: „Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm”, ETSI document ES 202 050 v1.1.3 (2003-11), Nov. 2003.

[6] H.G. Hirsch, D. Pearce: “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions”, ISCA workshop ASR2000, Paris, Frankreich, 2000

[7] H.G. Hirsch: „Automatic speech recognition in adverse acoustic conditions”, in “Advances in Digital Speech Transmission”, John Wiley and sons, 2008

[8] H.-G. Hirsch, C. Ehrlicher: “Noise estimation techniques für robust speech recognition“, ICASSP, 1995

### *Anhang – Adaptive Filterung mit cepstraler Glättung*

Im Folgenden findet sich eine mathematische Beschreibung der adaptiven Filterung mit cepstraler Glättung, wie sie in dem entwickelten Merkmalsextraktionsverfahren eingesetzt wird.

- Höhenanhebung: $x_{pre}(n) = x(n) - 0{,}97 \cdot x(n-1)$
- Betrachtung von Signalabschnitten mit jeweils 200 Abtastwerten, die mit einem Hamming Fenster gewichtet werden. Verschiebung des Analysefensters um 80 Abtastwerte
- Zero padding auf 256 Werte
- Als Ergebnis der Transformation von jeweils 256 Werten mit einer FFT erhält man das komplexe Spektrum

  $Y_k(\ell)$ *mit* $k = index\ of\ FFT\ bin\ und\ \ell = frame\ index;\quad 0 \le k \le 128$
- Mit Hilfe einer Funktion zur Schätzung des Störspektrums wird fortlaufend das Leistungsdichtespektrum $P_n(k)$ einer als stationär angenommenen Hintergrundstörung geschätzt.
- Es wird das aposteriori SNR bestimmt: $\gamma_k(\ell) = \max\left\{\frac{|Y_k(\ell)|^2}{P_n(k)}, 1\right\}$
- Aus dem aposteriori SNR wird ein apriori SNR geschätzt: $\hat{\xi}_k^{ml}(\ell) = \max\left\{\gamma_k(\ell) - 1, \xi_{\min}^{ml}\right\}$ *mit* $10\log_{10}\xi_{\min}^{ml} = -25dB$
- Mit Hilfe des apriori SNR wird das Leistungsdichtespektrum des ungestörten Sprachsignals geschätzt: $\hat{P}_S^{ml}(k,\ell) = \hat{\xi}_k^{ml}(\ell) \cdot P_n(k)$

- Das Leistungsdichtespektrum wird in den Cepstralbereich transformiert: $\hat{P}_S^{ceps}(q,\ell) = IDFT\{\ln(\hat{P}_S^{ml}(k,\ell))\}$ *mit* $q = index\ of\ cepstral\ bin$
- Das Cepstrum wird mit einem adaptiven Faktor $\sigma(q,\ell)$, dessen Wert sich individuell für jeden Cepstral-Koeffizienten und für jeden Signalabschnitt $\ell$ verändern kann, geglättet: $\overline{P}_S^{ceps}(q,\ell) = \sigma(q,\ell)\cdot\overline{P}_S^{ceps}(q,\ell-1) + (1-\sigma(q,\ell))\cdot\hat{P}_S^{ceps}(q,\ell)$
- Zur Herleitung der adaptiven Faktoren $\sigma(q,\ell)$ wird zunächst ein von dem Zeitindex $\ell$ unabhängiger Faktor betrachtet: $\overline{\sigma}^{const}(q) = \left\{\begin{array}{lll} 0{,}2 & für & q \in \{0,1,2,3\} \\ 0{,}99 & für & q \in \{4,\ldots,128\} \end{array}\right\}$
- Damit werden die Cepstral-Koeffizienten mit q>3 stark geglättet. Am Ausgang des Glättungsfilters findet man im Wesentlichen nur noch den Gleichanteil des jeweiligen Cepstral-Koeffizienten. Die niedrigen Cepstral-Koeffizienten mit q<4, die die Einhüllende des Leistungsdichtespektrums beschreiben, werden nur geringfügig geglättet.
- Im Folgenden werden die Cepstral-Koeffizienten, die der Grundfrequenz stimmhafter sprachlicher Abschnitte entsprechen, im Bereich stimmhafter Abschnitte ebenfalls nur geringfügig geglättet: $\sigma(q,\ell) = \left\{\begin{array}{lll} 0{,}2 & für & q \in Q_{pitch} \\ \overline{\sigma}(q,\ell) & für\ alle\ anderen\ Werte\ von\ q & \end{array}\right\}$
- $Q_{pitch}$ beschreibt die Menge aller Cepstral-Indices, die der Grundfrequenz zugeordnet werden. Diese Menge wird mit Hilfe einer Funktion zur Bestimmung der Grundfrequenz bestimmt, die in einem separaten Abschnitt des Anhangs beschrieben wird.
- Um die geringfügigere Glättung der Cepstral-Koeffizienten, die der Grundfrequenz zugeordnet werden, am Ende eines stimmhaften Abschnitts wieder langsam an die starke Glättung heranzuführen, wird die folgende, rekursive Anpassung der Glättungsfaktoren über der Zeit vorgenommen: $\overline{\sigma}(q,\ell) = \varphi\cdot\sigma(q,\ell-1) + (1-\varphi)\cdot\overline{\sigma}^{const}(q)$ *mit* $\varphi = 0{,}96$
- Aus dem geglätteten Cepstrum $\overline{P}_S^{ceps}(q,\ell)$ wird mit Hilfe einer DFT das zugehörige lineare Leistungsdichtespektrum bestimmt:

  $\hat{P}_S(q,\ell) = \exp\left(DFT\{\overline{P}_S^{ceps}(q,\ell)\} + \kappa\right)$ *mit* $\kappa = 0{,}3$

  Die additive Konstante $\kappa$ dient der Bias-Korrektur, die auf Grund der Glättung der Cepstren notwendig ist.

- Aus dem Leistungsdichtespektrum wird wiederum ein apriori SNR geschätzt:

$$\hat{\xi}_k^{ct}(\ell) = \max\left\{\frac{\hat{P}_s(k,\ell)}{P_n(k)}, \xi_{\min}\right\} \quad mit\ 10\log_{10}\xi_{\min} = -25dB$$

- Aus dem a priori SNR, das nach der cepstralen Glättung bestimmt wurde, wird mit Hilfe eines Ansatzes zur Schätzung der logarithmierten, spektralen Amplitude (LSA) eine Filterfunktion bestimmt:

$$G^{LSA}\left(\hat{\xi}_k^{ct}, \gamma_k\right) = \frac{\hat{\xi}_k^{ct}}{1+\hat{\xi}_k^{ct}} \cdot e^{0{,}5\cdot \exp\mathrm{int}\left(\frac{\hat{\xi}_k^{ct}}{1+\hat{\xi}_k^{ct}}\cdot\gamma_k\right)}$$

- Die Werte der Filterfunktion werden auf den Bereich zwischen einem minimalen Wert und 1 beschränkt:

$$G_k(\ell) = \max\left\{G_{\min}, \min\left\{G^{LSA}\left(\hat{\xi}_k^{ct}(\ell), \gamma_k\right), G_{\max}\right\}\right\}$$
$$mit \quad 20\cdot\log_{10} G_{\min} = -30dB \quad und \quad G_{\max} = 1$$

- Mit Hilfe der Filterfunktion wird aus dem Spektrum des gestörten Signals das Spektrum des ungestörten Sprachsignals geschätzt: $\hat{S}_k(\ell) = G_k(\ell)\cdot Y_k(\ell)$
- Aus dem geschätzten Spektrum des ungestörten Sprachsignals wird gemäß Bild 1 ein Wert für die in dem Signalabschnitt enthaltene Energie bestimmt: $\log E = \ln\left(\sum_{k=5}^{128}\left|\hat{S}_k(\ell)\right|^2\right)$

  Zur Berechnung der Energie werden die Spektralkomponenten unterhalb von ca. 150 Hz nicht berücksichtigt, da Sprache dort keine wesentlichen Energieanteile besitzt, jedoch viele Störgeräusche in diesem Bereich einen Teil ihrer Energie haben.
- Aus dem geschätzten Spektrum des ungestörten Sprachsignals wird zudem gemäß Bild 1 das Mel-Spektrum berechnet. Die logarithmierten Mel Spektralwerte werden mit einer DCT in den Cepstralbereich transformiert.

### *Anhang – Bestimmung der Grundfrequenz*

Zur Realisierung der cepstralen Glättung wird eine Schätzung der Grundfrequenz benötigt. Damit wird die Menge $Q_{pitch}$, die die Indices, die der Grundfrequenz zugeordnet werden, festgelegt. Das Verfahren zur Schätzung der Grundfrequenz wird nachstehend kurz erläutert und mathematisch beschrieben. Mit diesem Verfahren konnten geringfügig bessere Erkennungsergebnisse erzielt werden als mit dem in [1] beschriebenen Verfahren.

- Zur cepstralen Glättung wurde bereits aus dem geschätzten Leistungsdichtespektrum $\hat{P}_S^{ml}(k,\ell)$ der ungestörten Sprache das Cepstrum bestimmt:
  $\hat{P}_S^{ceps}(q,\ell) = IDFT\{\ln(\hat{P}_S^{ml}(k,\ell))\}$ *mit q = index of cepstral bin*
- Da die Maxima im DFT Spektrum, die bei einem stimmhaften Sprachsignalabschnitt im Abstand der Grundfrequenz auftreten, hauptsächlich im niedrigeren Frequenzbereich hervortreten, wird ein weiteres Cepstrum aus den DFT Komponenten unterhalb von 2 kHz bestimmt. Dazu werden die DFT Komponenten, die bis 2 kHz auftreten, in umgekehrter Reihenfolge im Bereich von 2 bis 4 kHz angeordnet. Die Anordnung der DFT Komponenten im Bereich von 0 bis 4 kHz wird dann noch einmal wiederholt im Bereich von 4 bis 8 kHz angeordnet:

$$\hat{P}_{Slow} = \{\hat{P}_S^{ml}(k=0), \hat{P}_S^{ml}(k=1);\ldots,\hat{P}_S^{ml}(k=64); \hat{P}_S^{ml}(k=63),\ldots,\hat{P}_S^{ml}(k=1),$$
$$\hat{P}_S^{ml}(k=0), \hat{P}_S^{ml}(k=1);\ldots,\hat{P}_S^{ml}(k=64); \hat{P}_S^{ml}(k=63),\ldots,\hat{P}_S^{ml}(k=1)\}$$
$$\hat{P}_{Slow}^{ceps}(q,\ell) = IDFT\{\ln(\hat{P}_{Slow}(k,\ell))\} \quad mit\ q = index\ of\ cepstral\ bin$$

- Es wird das Maximum des Cepstrums $\hat{P}_{Slow}^{ceps}(q,\ell)$ und der Index $q_{pitch}$ im Bereich der Indices $25 \le q \le 113$ bestimmt, was dem Grundfrequenzbereich von ca. 70 Hz bis 320 Hz entspricht.
- Ist $\hat{P}_S^{ceps}(0,\ell) \ge 1$ (bestimmte Mindestenergie) und $\hat{P}_S^{ceps}(1,\ell) \ge 0$ (mehr Energie im Bereich von 0 bis 2 kHz als im Bereich von 2 bis 4 kHz), so wird das ermittelte Maximum mit einem vom Index q abhängigen Schwellwert verglichen. Wird der Schwellwert überschritten, so geht man davon aus, dass es sich um einen stimmhaften Laut handelt. Der zugehörige Matlab Code:
  [mxval, kpitch] = max(ceps_low(26:114);
  thres = (0.4 – 0.25/89 * kpitch) * 2;
- Im Fall eines stimmhaften Lauts wird dann nochmals das Maximum $q_{pitch\_total}$ im Cepstrum $\hat{P}_S^{ceps}(q,\ell)$ des ungefilterten Spektrums im Bereich $q_{pitch} - 2 \le q \le q_{pitch} + 2$ bestimmt.
- Die Menge $Q_{pitch}$ besteht aus den 3 Werten $Q_{pitch} = \{q_{pitch\_total} - 1, q_{pitch\_total}, q_{pitch\_total} + 1\}$

### *Anhang – Blinde Frequenzgangkompensation*

Die Wichtung eines Betragsspektrums |S(f)| mit einer unbekannten Übertragungsfunktion |H(f)| tritt nach einer Logarithmierung der Spektren bzw. auch nach Transformation der logarithmierten Spektren in den Cepstralbereich als additiver Anteil auf:

$$|X(k)| = |S(k)| \cdot |H(k)|$$
$$\log|X(k)| = \log|S(k)| + \log|H(k)|$$
$$X^{ceps}(q) = S^{ceps}(q) + H^{ceps}(q)$$

In [4] wird ein Ansatz vorgestellt, bei dem durch eine fortlaufend über der Zeit stattfindende Differenzbildung des aus der Analyse eines Sprachabschnitts resultierenden Cepstrums und den Cepstralwerten, die ein mittleres Sprachspektrum definieren, eine Schätzung des unbekannten Frequenzgangs vorgenommen wird:

$$\hat{H}^{ceps}(q,\ell) = \hat{H}^{ceps}(q,\ell-1) + \alpha \cdot \left[X^{ceps}(q,\ell-1) - S^{ceps}_{average}(q)\right]$$

Die Werte $S^{ceps}_{average}(q)$ *mit* $1 \le q \le 12$ zur Festlegung des mittleren Sprachspektrums und der Faktor $\alpha$ sind in [4] definiert.

Mit den geschätzten Cepstralkoeffizienten der Übertragungsfunktion wird eine Schätzung der Cepstralkoeffizienten des Sprachspektrums vorgenommen:

$$\hat{S}^{ceps}(q,\ell+1) = X^{ceps}(q,\ell) + \hat{H}^{ceps}(q,\ell)$$

### *Anhang – Weitere Erkennungsergebnisse für eine robuste Merkmalsextraktion in Kombination mit einer Adaption der HMMs auf eine stationäre Hintergrundstörung*

Im Fokus dieses Projekts steht die Betrachtung einer Spracheingabe im Freisprechmodus in einer gestörten Umgebung. Da sich bei den Untersuchungen zur Kombination einer robusten Merkmalsextraktion mit einer Adaption der HMMs recht positive Ergebnisse für die Kombination einer robusten Merkmalsextraktion und einer Adaption der HMMs auf eine stationäre Störung einstellten, wurden einige weitere Erkennungsexperimente mit Sprachdaten, bei denen die additive Überlagerung einer Hintergrundstörung den dominierenden Effekt darstellt, durchgeführt. In den Tabellen 12 bis 15 werden die Fehlerraten bei Überlagerung eines Störgeräuschs im Auto bzw. typischer Störgeräusche in einer räumlichen Umgebung aufgeführt. In allen Fällen lässt sich feststellen, dass die zusätzliche Adaption der HMMs auf das Störgeräusch eine Verbesserung der Erkennung bewirkt. Der Vergleich mit den Ergebnissen bei Einsatz der von ETSI standardisierten

Merkmalsextraktion zeigt, dass mit dem betrachteten Ansatz in den meisten Fällen eine bessere Erkennung gewährleistet werden kann.

| | | SNR/dB | | | |
|---|---|---|---|---|---|
| Analyseverfahren | Adaption | 15 | 10 | 5 | 0 |
| ETSI-2 | | 1,35 % | 2,52 % | 5,94 % | 16,14 % |
| HGH-Robust | | 1,23 % | 2,24 % | 5,87 % | 17,45 % |
| HGH-Robust | auf Störung | 1,36 % | 2,33 % | 5,75 % | 15,95 % |

Tabelle 13: „Aurora-5“ Wortfehlerraten für die gestörte Umgebung im Auto (ohne Freisprechen)

| | | SNR/dB | | | |
|---|---|---|---|---|---|
| Analyseverfahren | Adaption | 15 | 10 | 5 | 0 |
| ETSI-2 | | 4,17 % | 7,62 % | 15,42 % | 33,48 % |
| HGH-Robust | | 3,00 % | 5,99 % | 14,57 % | 35,66 % |
| HGH-Robust | auf Störung | 3,04 % | 5,66 % | 12,83 % | 31,39 % |

Tabelle 14: Wortfehlerraten für „Aurora-5“ in der akustischen Umgebung „car noise handsfree“

| | | SNR/dB | | |
|---|---|---|---|---|
| Analyseverfahren | Adaption | 15 | 10 | 5 |
| ETSI-2 | | 2,48 % | 5,38 % | 13,10 % |
| HGH-Robust | | 2,92 % | 6,36 % | 15,75 % |
| HGH-Robust | auf Störung | 2,57 % | 5,50 % | 13,03 % |

Tabelle 15: „Aurora-5“ Wortfehlerraten für eine gestörte räumliche Umgebung (ohne Freisprechen)