MULTIQUELLENTRAINING: CHANCEN FÜR KLEINE TRAININGSMENGEN IN DER AUTOMATISCHEN SPRACHERKENNUNG

Matthias Wölfel

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

wolfel@ira.uka.de

Kurzfassung: Eine große Anzahl an Faktoren beeinflussen die Akustik von Sprache, z.B. Sprachstil, Geschwindigkeit, Dialekt, Muttersprache, Hintergrundgeräusche und Hall. Ein Ungleichgewicht dieser Faktoren zwischen Trainings- und Testmaterial eines automatischen Spracherkenners mindert dessen Leistung erheblich. Deshalb werden Spracherkenner auf eine bestimmte Aufgabe trainiert, in dem man möglichst viele Daten mit Eigenschaften die den späteren Einsatzbedingungen entsprechen sammelt und transskribiert. Diese Arbeit ist sehr kosten- und zeitintensiv und führt oft, wie sich aus dem Schlagwort ”There’s no data like more data!” folgern lässt, nicht zu einer ausreichenden Anzahl an Daten um gute Ergebnisse zu erzielen. In unserer Arbeit zeigen und untersuchen wir verschiedene Verfahren um diese Lücke zu schliessen. Dies wird erreicht indem Trainingsmaterial aus einer anderen Quelle die im Sprachstil und akustischer Umgebung stark von dem Testmaterial abweichen kann, mit ein paar Stunden Adaptionsmaterial, das aus gleicher akustischer Umgebung stammt wie das Testmaterial, geschickt kombiniert wird. Hierfür untersuchten wir die folgenden Ansätze:

- Training des akustischen Modells des Spracherkenners mit Sprachmaterial das sich in Sprachstil und akustischer Umgebung stark vom zu erkennenden Sprachmaterial unterscheidet und anschliessende überwachte Adaption mit aus gleicher Quelle stammendem Sprachmaterial wie die zu testende Sprache.
- Berechnung der linearen Diskriminanzanalyse die auf dem Adaptionsmaterial und nicht wie üblich auf dem Trainingmaterial basiert.
- Einführung einer sprecherabhängigen Modellordnung der Minimum Variance Distortionless Response Einhüllenden.
- Unüberwachte sprecherabhängige Adaption des akustischen Modells anhand von Hypothesen die aus einem anderen akustischen Modell stammen.

Alle Untersuchungen wurden mit dem automatischen Spracherkenner *Janus Recognition Toolkit* (JRTk) der Interactive Systems Laboratories (Universität Karlsruhe, Deutschland und Carnegie Mellon University, USA) durchgeführt, tabellarisch zusammengestellt und werden diskutiert.

1 Einleitung

Es gibt eine Vielzahl an Faktoren, die die Akustik von Sprache beeinflussen und es dadurch dem Menschen als auch der Maschine erschweren die gesprochenen Wörter zu verstehen. Es seien hier einige Beispiele genannt: Sprachstil (monoton, emotional oder hyperartikuliert), Geschwindigkeit (schnell, langsam oder sogar variierend), Dialekt (in diesem Fall kann sogar das verwendete Vokabular variieren), Muttersprache (bei Nichtmuttersprachlern geht die Muttersprache sehr stark in die Aussprache mit ein), Stimmhöhe (Frau, Mann oder Kind), Hintergrundgeräusche (z.B. Klimaanlage oder andere Gespräche) sowie Hall und Echo. Ein Ungleichgewicht der soeben genannten Faktoren zwischen Trainings- und Testmaterial eines automatischen Spracherkenners kann zu einer erheblichen Minderung an Wortgenauigkeit führen.

In der Regel versucht man eine bestmögliche Erkennung zu erzielen indem man ein Spracherkennersystem auf eine bestimmte Aufgabe trainiert, in dem man möglichst viele Daten mit Eigenschaften die den späteren Einsatzbedingungen entsprechen sammelt und transskribiert. Das Sammeln von entsprechenden Daten ist nicht nur sehr kosten- und zeitintensiv, ja sogar teilweise unmöglich, da sich die akustischen Bedingungen sehr schnell ändern können. Dies führt oft, wie sich aus dem Schlagwort ”There’s no data like more data!” folgern lässt, nicht zu einer ausreichenden Anzahl an Daten um gute Ergebnisse zu erzielen.

Im Folgenden zeigen und untersuchen wir verschiedene Verfahren, die wir im Kapitel *Theoretischer Hintergrund* kurz erklären um sie dann im Kapitel *Experimente & Diskussion* experimentell zu untersuchen und zu diskutieren.

2 Theoretischer Hintergrund

In diesem Kapitel beschäftigen wir uns zuerst mit der Anpassung des Sprechers und des Kanals an die akustischen Modelle des Spracherkenners. Hierzu betrachten wir zunächst das Maximum Likelihood Kriterium um anschliessend die Vokaltraktlängennormierung und die Maximum Likelihood Linear Regression zu erklären, die beide die Likelihood als Optimierungskriterium verwenden um die freien Parameter einzustellen. Danach wird kurz auf die lineare Diskriminanzanalyse eingegangen und begründet warum es sinvoll ist die lineare Diskriminanzanalyse auf Basis des Adaptionsmaterials zu berechnen. Des weiteren wollen wir kurz die Minimum Variance Distortionless Response Einhüllenden und deren Variationen des Skalierens und Mel-Frequenzanpassens vorstellen und diese um eine sprecherabhängige Modellordnung erweitern. Den theoretischen Hintergrund wollen wir abrunden in dem wir auf die Kreuzadaption eingehen, bei der Hypothesen zur Adaption des akustischen Modells verwendet werden, die aus einem anderen akustischen Modell stammen.

2.1 Vokaltraktlängennormierung und Maximum Likelihood Linear Regression

Das Ziel von Adaptionsverfahren in der automatischen Spracherkennung besteht darin, bestehende Modelle λ (z.B. Hidden Marcov Modelle) durch Adaption an eine neue Situation (z.B. Sprecher und Mikrophon) anzupassen und somit die Wortfehlerrate zu reduzieren. Um dieses Ziel zu erreichen wird ein zuverlässiges Kriterium gesucht, mit dem entschieden werden kann, ob sich eine Verbesserung erzielen läßt. Ein solches Kriterium ist die *Maximum Likelihood* (ML). Hier werden die Parameter im Eigenschafts- oder Modellraum so an das akustische

Modell X des Testsprechers angepaßt, bis die Likelihood (die Wahrscheinlichkeit und somit das Ähnlichkeitsmaß) $P(X|\lambda)$ anhand der zum akustischen Modell gefundenen Hypothese W (ein Wort oder eine Wortfolge) maximiert ist:

$$\text{Eigenschaftsraum}: \quad \alpha_{\text{ML}} = \arg\max_{\alpha} P(X(\alpha)|W,\lambda)$$

$$\text{Modellraum}: \quad \alpha_{\text{ML}} = \arg\max_{\alpha} P(X|W,\lambda(\alpha))$$

Als Beispiel sei hier die *Vokaltracklängennormierung* (VTLN) genannt, bei der die optimale Verschiebung des Spektrums im Frequenzbereich durch das ML Kriterium gefunden wird.

Stehen nur wenige Sprachdaten aus einer von der trainierten abweichenden akustischer Umgebung zur Verfügung, können nicht alle Parameter des akustischen Modells nach dem ML Kriterium zuverlässig bestimmt werden. Um dennoch eine zuverlässige Schätzung der Modelle zu ermöglichen verwendet man die *Maximum Likelihood Linear Regression* (MLLR) bei der die Anzahl an freien Parametern verringert wird indem man die Modelle clustert. Hierdurch erhöht sich die für die einzelnen Parameter zu Verfügung stehende Adaptionsmenge, wodurch eine zuverlässigere Adaption der jetzt nur noch geringen Anzahl von Parametern möglich ist. Wie schon bei der VTLN wird auch hier das ML Verfahren als Optimierungskriterium verwendet, aber hier durch Anpassung der *Hidden Marcov Modell* (HMM)-Parameter durch Multiplikation der Mittelwertvektoren μ_{s} mit der Regressionsmatrix M_s [9]:

$$\mu_{\text{MLLR,s}} = M_{\text{s}} \cdot \mu_{\text{s}}$$

$$M_{\text{s,ML}} = \arg\max_{M_{\text{s}}} P(X|W,\lambda(\mu_{\text{MLLR,s}}))$$

2.2 Lineare Diskriminanzanalyse

Um die Anzahl der Eigenschaften, in einem Standardspracherkenner die Kepstralkoeffizienten, im Modell gering zu halten und nicht unnötig Redundanz zu modellieren ist es sinnvoll die Dimension des akustischen Modells zu reduzieren. Ein geeignetes Verfahren um dies zu erreichen ist die *lineare Diskriminanzanalyse* (LDA). Ihr Ansatz basiert auf einer linearen Transformation $y = Ax$, der den Vektor x mit Dimension l auf einen Vektor y der Dimension $m < l$ reduziert. Hierbei sollen die Werte von A derart bestimmt werden, dass die durch die Lernstichprobe gegebene Zerlegung der x-Werte durch die y-Werte 'möglichst gut' wiedergegeben werden. Dies ist genau dann der Fall wenn die durch die lineare Transformation erhaltenen Gruppenmittelwerte $\overline{y}_i$ möglichst weit voneinander entfernt sind.

Der Abstand der Gruppenmittelwerte sollte sich im besten Fall auf das Testmaterial beziehen, da durch eine Erhöhung der Distanz der Gruppenmittelwerte des Testmaterials eine bessere Trennbarkeit erreicht werden kann. Deshalb berechnen wir die LDA nicht wie allgemein üblich auf Basis von Trainingsmaterial sondern auf Basis des Adaptionsmaterials das dem zu testenden Material stärker ähnelt.

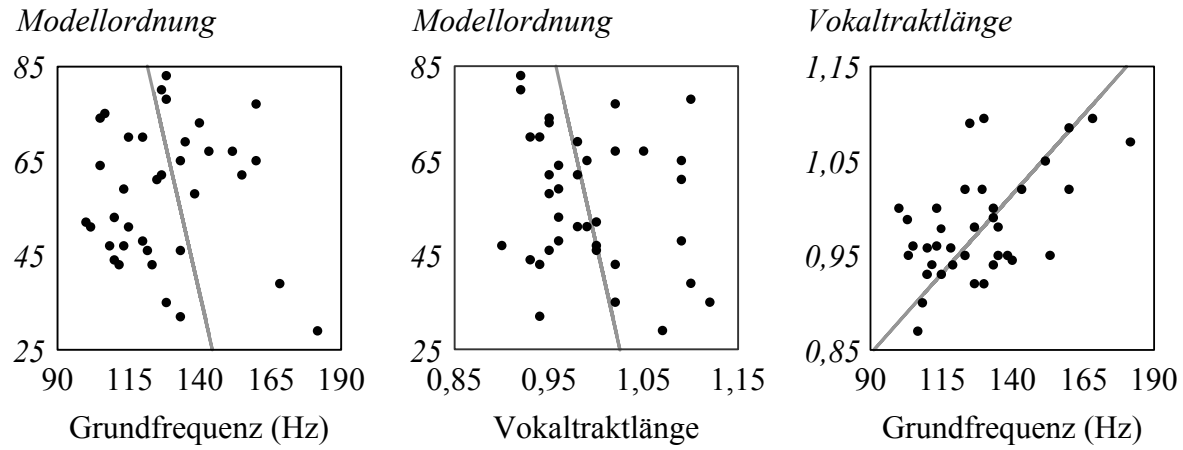

Abbildung 1 - Zusammenhänge zwischen der Modellordnung, Grundfrequenz und Vokaltraktlänge für die 39 TED Sprecher .

2.3 Sprecherabhängige Modellordnung der skalierten und Mel-Frequenzangepaßten Minimum Variance Distortionless Response Einhüllenden

Anstelle der in der Spracherkennung weit verbreiteten Mel-Frequenz Kepstralkoeffizienten oder der linearen Prädiktionskoeffizienten verwenden wir *Minimum Variance Distortionless Response*[1] (MVDR) Kepstralkoeffizienten die von Dharanipragada und Rao [3] zum ersten Mal in der Spracherkennung, in einer von uns abweichenden Form, eingesetzt wurden. Eine gute Übersicht zur MVDR Einhüllenden, ohne direkten Bezug auf Spracherkennung, ist die Veröffentlichung von Murthi und Rao [4]. In unserem System verwenden wir eine Mel-Freqenz adaptierte Form, die auf der bilinearen Transformation basiert, um die nichtlineare Frequenzauflösung ähnlich dem menschlichen auditiven System zu berücksichtigen, sowie eine Skalierung der höchsten Energie auf die höchste Energie der Fouriertransformation [5, 6, 7]. Um eine bessere Adaption an den Sprecher bzw. dessen Grundfrequenz zu erreichen, passen wir die Modellordnung an den Sprecher mit Hilfe des ML Kriteriums an [8]. Hierzu betrachten wir die Modellordnung N als freien Parameter und verändern diesen, pro Sprecher, so lange bis sich die Likelihood nicht weiter erhöht.

$$N_{\mathrm{ML}} = \arg\max_{N} P(X(N)|W,\lambda)$$

Bild 1 zeigt die Zusammenhänge zwischen der Modellordnung und Grundfrequenz sowie Vokaltraktlänge für die 39 TED Sprecher. Hierbei steht jeder Punkt für einen Sprecher und die graue Linie für die Regressionslinie. Diese wird berechnet indem man die Linie so einpasst, dass der quadratische Abstand zu den Punkten minimiert wird. Aus den Schaubildern wird ersichtlich, dass nur ein schwacher Zusammenhang der Modellordnung zu der Grundfrequenz sowie Vokaltraktlänge besteht. Zur Vollständigkeit ist auch der Zusammenhang zwischen der Grundfrequenz und Vokaltraktlänge gezeigt. Hier wird eine stärkere Abhängigkeit voneinander deutlich.

Mit einer Darstellung der schnellen Berechnung der Mel-Frequenzangepassten MVDR, Tabelle 1, und dem Verweis auf die zuvor referenzierte Literatur wollen wir es belassen.

[1] Die MVDR wurde zuerst von Capon [1] vorgestellt und ist auch bekannt als Maximum-Likelihood Methode.

1. **Berechnung der Mel-Frequenz LP Koeffizienten**
   Zur Berechnung der Mel-Frequenz LP Koeffizienten $\tilde{\mathbf{a}}$, der Ordnung N, gibt es verschiedene Möglichkeiten. Bei unseren Versuchen wurde ein Algorithmus wie von Matsumoto u.a. vorgeschlagen verwendet [2].

2. **Korrelation der Mel-Frequenz Vorhersagekoeffizienten**

$$\tilde{\mu}_k = \begin{cases} \sum_{i=0}^{N-k} (N+1-k-2i)\tilde{a}_i^{(N)} \tilde{a}_{i+k}^{*(N)} & : k = 0, \cdots, N \\ \tilde{\mu}_{-k}^* & : k = -N, \cdots, -1 \end{cases}$$

3. **Berechnung der Mel-Frequenz MVDR Einhüllenden**

$$\tilde{S}_{\mathrm{MVDR}}(\omega) = \frac{\epsilon}{\sum_{k=-N}^{N} \tilde{\mu}_k e^{-j\omega k}}$$

Tabelle 1 - Eine schnelle Methode um die Mel-Frequenzangepasste MVDR zu berechnen.

2.4 Kreuzadaption

Um das für die Spracherkennung akustische Modell zu adaptieren verwendet man üblicherweise Hypothesen, die mit dem selben akustische Modell generiert wurden. In dem hier beschriebenen Ansatz, den wir Kreuzadaption (*engl.* cross adaptation) nennen wollen, verwenden wir Hypothesen die mit einem anderen akustischen Modell erzeugt wurden.

3 Experimente & Diskussion

Die unten aufgeführten Spracherkennerexperimente wurden mit dem *Janus Recognition Toolkit* (JRTk) durchgeführt, das gemeinsam von den Schwesterlaboratorien, den *Interactive Systems Laboratories*, an der Universität Karlsruhe (TH), Deutschland und an der Carnegie Mellon University in Pittsburgh, Pennsylvania, USA entwickelt und gepflegt wird.

Alle Erkennerläufe wurden auf dem *Translanguage English Database* (TED) Korpus [10] durchgeführt, vertrieben vom *Linguistic Data Consortium* (LDC). Hierbei handelt es sich um kontinuierliche englische Sprache, aufgezeichnet auf der EUROSPEECH 1993. Da die meisten Vorträge nicht von englischen Muttersprachlern sind, und somit im Vergleich schwer (sowohl vom Menschen als auch automatisch) zu transskribieren sind, ist dieser Korpus auch bekannt als ”Terrible English Database”. Von den 39 transskribierten Sprechern wurden acht für die Erkennung extrahiert. Die restlichen 31 Sprecher standen als Adaptionsmaterial oder alternativ zum Training zur Verfügung. Als zusätzlicher Trainingskorpus kam *Broadcast News 96* (BN) [11] zum Einsatz.

Unser akustisches Modell verwendet 4.139 Kodebücher mit je 32 Gaussglocken. Von dem mit 16 kHz abgetasteten Signal werden alle 10 ms neue Merkmale berechnet. Hierfür zerlegen wir das kontinuierliche Sprachsignal durch ein 20 ms Hamming Fenster und berechnen die skalierte Mel-Frquenz MVDR Einhüllende. Danach werden 13 Kepstralkoeffizienten, die durch kepstrale Mittelwertssubtraktion normalisiert wurden, zusammen mit ihren drei linken und rechten Nachbarn durch eine diskrete Kosinustransformation berechnet. Die Merkmale werden zu ihrer endgültigen Anzahl, 40, durch eine LDA reduziert.

Das verwendete 3-Gramm Sprachmodell berücksichtigt sowohl Veröffentlichungen von Konferenzen wie EUROSPEECH, ICSLP und ICASSP als auch die Vorträge der 31 Adaptionssprecher um Effekte der gesprochenen Sprache zu modellieren. Das Vokabular umfaßt 25.000 Wörter wobei im Schnitt 1,6 Aussprachevarianten berücksichtigt wurden. Die Rate der nicht im Vokabular enthaltenen Wörter beträgt ca. 0,3%.

In Tabelle 2 sind die Ergebnisse unserer Versuche zusammengestellt. Aus der ersten Spalte, bezeichnet als ”Training”, lässt sich das Trainingsmaterial entnehmen mit dem die akustischen Modelle des Spracherkenners trainiert wurden. Der Zusatz ”LDA auf 31 TED Sprechern” weist darauf hin, dass hier anstelle des BN Korpus die 31 TED Sprecher zur Berechnung der LDA verwendet wurden. Der zweiten Spalte, bezeichnet als ”Adaption (31 TED Sprecher)”, läßt sich entnehmen mit welchen Verfahren auf den 31 TED Sprecher adaptiert wurde. Hierbei handelt es sich um eine überwachte Adaption, das bedeutet, es wird anhand der Referenz (von Hand transskribierter Text) adaptiert. Die VTLN und Modellordnung wurden für jeden Sprecher individuell adaptiert, wohingegen die MLLR über alle 31 Sprecher berechnet wurden um auf den Kanal zu adaptieren. In der dritten Spalte ”Adaption (8 TED Testsprecher)” sind die unüberwachten Adaptionsverfahren genannt. Im Gegensatz zur Adaption der 31 Sprecher wurde hier die MLLR nicht mehr über alle Sprecher des Testsets gemeinsam berechnet, sondern getrennt für jeden einzelnen, um die Eigenschaften des jeweiligen Sprechers bei der Anpassung mitzuberücksichtigen. In der letzten Spalte ”WER” können die Wortfehlerraten entnommen werden.

Zunächst wollen wir die beiden Basissysteme mit MLLR und VTLN Adaption vergleichen. Obwohl sich das BN System stark vom Testset unterscheidet schneidet es mit 2% absoluter WER besser ab als das auf TED Sprecher trainierte System. Der Grund hierfür liegt in der geringen Datenmenge des TED Trainingsmaterials, ca. 8 Stunden, im Vergleich zu ca. 100 Stunden des BN Korpus. Wird eine überwachte MLLR und VTLN Adaption der 31 TED Sprecher hinzugefügt verringert sich der Fehler um 3,1% absolut. Durch Berechnung der LDA, auf Basis des Adaptionsmaterials, kann die Wortfehlerrate weiter verringert werden, da wie bereits im theoretischen Teil erwähnt, die Klassifizierbarkeit der TED Testsprecher weiter erhöht wird. Eine Modellordnungsanpassung der Testsprecher führt zu keiner statistisch signifikanten Verbesserung. Erst durch zusätzliche Modellordnungsanpassung der Adaptionssprecher verringert sich die WER um absolut 1%. Durch Kreuzadaption kann keine statistisch signifikante Verbesserung erzielt werden. Zuerst scheint das Ergebnis, dessen Modelle auf den 31 TED Sprechern trainiert wurden, überraschend, da es statistisch gesehen genau so gut abschneidet wie das BN System. Allerdings müssen wir hier berücksichtigen, dass die Daten auf Hypothesen mit einer WER von 37,4% trainiert wurden und somit das Ergebnis nicht verbessert werden konnte. Die Hypothesen des kreuzadaptierten Modells verwenden die Hypothesen, die mit dem TED System generiert wurden, d.h. mit einer WER von 44,4%. Im direkten Vergleich zu nicht kreuzangepassten Modellen schneidet das kreuzangepasste BN System um 0,2% besser ab, somit kann nicht von einer signifikanten Verbesserung gesprochen werden.

Als Vergleichswert zur Mel-Frequenz MVDR Spektralberechnung wird eine Fourier Spektralberechnung verwendet (nicht in der Tabelle aufgelistet). Auf dem Broadcast News 96 Task, mit überwachtem MLLR und VTLN Adaption auf Basis der 31 TED Adaptionssprecher und unüberwachter MLLR und VTLN Adaption der Testsprecher wird eine WER von 39,8% erreicht, d.h. die Fouriertransformierte schneidet hier mit 0,7% absolut, bzw. 1,8% relativ, schlechter ab als die Mel-Frequenz MVDR Spektralberechnung.

Aus der vorangegangen Analyse unserer Versuchsergebnisse kann geschlossen werden, dass sich durch Mischen, genauer gesagt durch überwachte Adaption von Quellen, die dem zu

Training	***Adaption*** ***(31 TED Sprecher)***	***Adaption*** ***(8 TED Testsprecher)***	***WER***
Broadcast News 96 LDA auf 31 TED Sprechern	MLLR, Modellordnung, VTLN	MLLR, Modellordnung, VTLN, Kreuzadaption	37,2%
31 TED Sprecher		MLLR, VTLN, Kreuzadaption	37,3%
Broadcast News 96 LDA auf 31 TED Sprechern	MLLR, Modellordnung, VTLN	MLLR, Modellordnung, VTLN	37,4%
Broadcast News 96 LDA auf 31 TED Sprechern	MLLR, VTLN	MLLR, Modellordnung, VTLN	38,3%
Broadcast News 96 LDA auf 31 TED Sprechern	MLLR, VTLN	MLLR, VTLN	38,4%
Broadcast News 96	MLLR, VTLN	MLLR, VTLN	39,1%
Broadcast News 96		MLLR, VTLN	42,2%
Broadcast News 96			64,5%
31 TED Sprecher		MLLR, VTLN	44,4%
31 TED Sprecher			51,3%

Tabelle 2 - Ergebnisse, in Wortfehlerrate, der im Text beschriebenen Versuche

transskribierenden Material ähnlich sind, die WER eines Spracherkenners verbessern kann. Des weiteren konnte gezeigt werden, dass es zu einer Verbesserung führen kann die LDA auf Daten zu berechnen die dem Testset ähneln, selbst wenn die zur Verfügung stehende Datenmenge im Vergleich zur Trainingsmenge gering ist. Durch die Einführung einer variablen Modellordnung konnte eine weitere Verbesserung erzielt werden. Auf den von uns untersuchten Daten erwies sich die Verbesserung durch Kreuzadaption als nicht signifikant.

Aus einer Analyse unserer Versuchergebnisse, in Tabelle 2 zusammengestellt, sehen wir wie sich durch Mischen von Trainingsmaterial aus verschiedenen Quellen und deren Adaption die Genauigkeit eines Spracherkenners verbessern läßt.

Um die Spracherkennung weiter zu verbessern kann zusätzlich berücksichtigt werden, dass die Sprecher unterschiedliche Muttersprachen sprechen und somit eine Adaption auf dieser Basis, d.h. eine Auswahl nach Sprache oder Sprachgruppe oder nach dem Maximum Likelihood Kriterium, möglich ist. Bei ersten Versuchen diesbezüglich konnte die WER einzelner Sprecher, sowohl bei der handsortierten Auswahl nach Sprachgruppen als auch bei einem unüberwachten Maximum Likelihood Ansatz, um bis zu 2% der absoluten Fehlerrate verbessert werden. Leider gab es bei anderen Sprechern einen Verlust an Wortgenauigkeit, so dass die Gesamtleistung des Spracherkenners nicht verbessert werden konnte. Hier muss ein Kriterium gefunden werden das es ermöglicht die Sprecher so zu selektieren, dass bei allen Sprechern ein Gewinn an Wortgenauigkeit eintritt.

Literatur

[1] Capon, J. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE, vol. 57:pp. 1408–1418, August 1969.*

[2] Matsumoto, H. und Moroto, M. Evaluation of Mel-LPC cepstrum in a large vocabulary continuous speech recognition. *IEEE International Conference on Acoustic, Speech, and Signal Processing 2001.*

[3] Dharanipragada, S. und Rao, B.D. MVDR based feature extraction for robust speech recognition *IEEE International Conference on Acoustic, Speech, and Signal Processing 2001.*

[4] Murthi, M.N. und Rao, B.D. All-pole modeling of speech based on the minimum variance distortionless response spectrum. *IEEE Transactions on Speech and Audio Processing, 2000*

[5] Wölfel, M. Mel-Frequenzanpassung der Minimum Varianz Distortionless Response Einhüllenden. *Eurospeech 2003.*

[6] Wölfel, M.; McDonough, J.W. und Waibel, A. Minimum variance distortionless response on a warped frequency scale. *Eurospeech 2003.*

[7] Wölfel, M.; McDonough, J.W. und Waibel, A. Warping and Scaling of the Minimum Variance Distortionless Response. *Automatic Speech Recognition and Understanding 2003.*

[8] Wölfel, M. Speaker Dependent Model Order Selection of Spectral Envelopes. *International Conference on Speech and Language Processing 2003.*

[9] Leggetter, C. J.; Woodland, P. C. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language, pp. 171–185, 1995.*

[10] Linguistic Data Consortium (LDC) Translanguage English Database *www.ldc.upenn.edu/Catalog/LDC2002S04.html*

[11] Linguistic Data Consortium (LDC) English Broadcast News Speech (Hub-4) *www.ldc.upenn.edu/Catalog/ LDC97S44.html*