Vorwärtsmaskierung für cepstrum-basierte Spracherkennungssysteme

Sascha Wendt, Gernot A. Fink, Franz Kummert

Universität Bielefeld, Technische Fakultät,
Postfach 100131, 33501 Bielefeld
email: {swendt,gernot,franz}@techfak.uni-bielefeld.de

Zusammenfassung

In der automatischen Sprachverarbeitung sind derzeit MFCC (Mel-frequency cepstral coefficients) oder LPCC (linear predictive cepstral coefficients) als Merkmale sehr verbreitet. Die Berechnung dieser Merkmale orientiert sich dabei i.a. nur rudimentär an der menschlichen Sprachwahrnehmung. Geht man jedoch davon aus, daß die Repräsentation von Sprache beim Menschen eine optimale Repräsentation darstellt, könnte eine stärkere Berücksichtigung von menschlichen Gehöreigenschaften zu einer Verbesserung von automatischen Spracherkennungssystemen führen. In diesem Beitrag wird ein Spracherkennungssystem mit einem von Strope und Alwan [1] vorgestellten Modell zur Berechnung des Effektes der Vorwärtsmaskierung erweitert, wobei einige Änderungen in der Anwendung des Modells vorgenommen wurden. Das Modell, das sich an der akustischen Wahrnehmung des Menschen orientiert, läßt sich sehr einfach in einem auf MFCC-Merkmalen beruhenden Spracherkennungssystem einsetzen, da nur wenige Anpassungen in der Merkmalsberechnung notwendig sind, und führt vor allem in geräuschvollen Aufnahmeumgebungen zu einer deutlichen Abnahme der Wortfehlerrate.

1. Einleitung

Aktuelle automatische Spracherkennungssysteme sind in ihrer Erkennungsleistung trotz intensiver Forschung noch immer weit von der Leistung der menschlichen Spracherkennung entfernt. Ein Grund dafür liegt in der akustischen Wahrnehmung des Menschen. Das menschliche Gehör verfügt über Eigenschaften, die bei der menschlichen Spracherkennung eine wichtige Rolle spielen, in modernen automatischen Spracherkennungssystemen jedoch nicht oder nur teilweise berücksichtigt werden. Die Einbindung solcher Eigenschaften hat sich bereits in einigen Fällen als sehr erfolgreich erwiesen. So hat sich beispielsweise die gehörorientierte Verzerrung des Frequenzspektrums mittels einer Mel-Filterbank in den meisten Spracherkennungssystemen etabliert. Ein solches Beispiel läßt vermuten, daß durch die Berücksichtigung anderer bekannter Gehöreigenschaften weitere Verbesserungen in der automatischen Spracherkennung möglich sind.

[1]Die vorliegende Arbeit wurde im Rahmen des Sonderforschungsbereichs 360 "Situierte künstliche Kommunikatoren" von der Deutschen Forschungsgemeinschaft (DFG) gefördert.

In W. Hess and K. Stoeber, editors, Elektronische Sprachsignalverarbeitung, volume 22 of Studientexte zur Sprachkommunikation, pages 85-91, Bonn, 2001.

Bei dem Versuch, diese Gehöreigenschaften in automatischen Spracherkennungssystemen zu realisieren, lassen sich grob zwei verschiedene Ansätze unterscheiden: Zum einen werden Spracherkennungssysteme entwickelt, die sich bei der Merkmalsberechnung fast vollständig auf Gehörmodelle stützen (siehe z.B. [2, 3]). Auf der anderen Seite gibt es verschiedene Ansätze, etablierte Merkmalsberechnungen wie MFCC oder LPCC, die sich nur rudimentär an der menschlichen Sprachwahrnehmung orientieren, durch Berücksichtigung von bekannten Gehöreigenschaften, wie z.B. der Maskierung (siehe u.a. [1, 4]), zu erweitern.

Mit dem zweiten Ansatz beschäftigt sich dieser Beitrag. Es wird zunächst ein Modell zur Berechnung des Effektes der Vorwärtsmaskierung, das in [1] vorgestellt wurde, modifiziert. Mit diesem modifizierten Modell wird dann ein Spracherkennungssystem, das auf MFCC-Merkmalen beruht, erweitert. Das Modell stellt dabei ein relativ einfaches Verfahren dar und läßt sich ohne große Änderungen der Merkmalsberechnung einbinden.

Im folgenden werden zunächst der Effekt der Vorwärtsmaskierung sowie das Modell zur quantitativen Erfassung bzw. die Änderungen in der Modellanwendung erläutert. Nach der Anpassung der Merkmalsberechnung erfolgt die Beschreibung der Evaluierung und die Diskussion der Ergebnisse. Eine Zusammenfassung bildet den Abschluß dieses Beitrags.

2. Vorwärtsmaskierung

Der Effekt der Vorwärtsmaskierung bewirkt, daß ein Schall die Lautstärkeempfindung eines zeitlich nachfolgenden Schalls beim Menschen beeinflußt, indem der nachfolgende Schall gedämpft wahrgenommen wird.

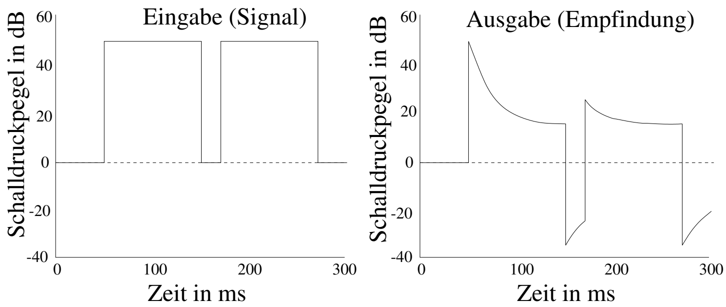

Abbildung 1: *Grafische Darstellung des Maskierungseffektes nach [1]: Die Empfindung des zweiten Signals wird durch das ersten Signal gedämpft.*

In Abbildung 1 ist dieser Effekt grafisch dargestellt: Zunächst erfolgt in Form eines Rechtecksignals eine Anregung über einen längeren Zeitraum hinweg, dann folgt ein zweites Rechtecksignal. Das erste Rechtecksignal wird vom Gehör adaptiert und dadurch in seiner Empfindung langsam gedämpft. Das zweite Rechtecksignal wird dann vom vorangegangenen Signal maskiert. Dies führt gleich zu Beginn des zweiten Signals, das ebenfalls langsam adaptiert wird, zu einer geringeren Empfindung.

Die Stärke der Maskierung hängt von verschiedenen Faktoren ab. Zum einen vom zeitlichen Abstand zwischen dem maskierenden und dem maskierten Signal: je größer der Abstand, desto geringer der Maskierungseffekt. Ebenso spielt die zeitliche Dauer des maskierenden Signals eine Rolle: je größer die Dauer ist, desto stärker werden nachfolgende Signale gedämpft. Weiterhin ist die Stärke des Maskierungseffektes von der Intensität des maskierenden Signals abhängig, denn je stärker das Signal ist, um so stärker wird ein nachfolgendes Signal maskiert. Diese Abhängigkeiten werden in den meisten Modellen durch verschiedene Parameter repräsentiert. Da die Maskierung auch von der Frequenz abhängt, wird i.d.R. für jede Frequenzgruppe jeweils ein eigener Parametersatz benötigt.

2.1. Modelle

In der aktuellen Forschung gibt es verschiedene Ansätze, um den Effekt der Vorwärtsmaskierung in automatischen Spracherkennungssystemen zu berücksichtigen.

In [3] and [5] werden z.B. sogenannte Rückkopplungsschleifen verwendet. Jede dieser Rückkopplungsschleifen hat ihre eigene Zeitkonstante und kann durch ein Signal aufgeladen werden bzw. wird entsprechend entladen, wenn kein Signal anliegt. Der Ladezustand dieser Rückkopplungsschleifen bestimmt dann, wie stark das aktuelle Signal maskiert wird. Dieses Modell wurde in einem Spracherkennungssystem verwendet, das sich fast vollständig auf Gehörmodelle stützt.

In [4] wird der Effekt der Vorwärtsmaskierung berücksichtigt, indem vom aktuellen Signal eine gewichtete Summe der vorangegangenen Signale subtrahiert wird. Dieses Modell wurde in einem Spracherkennungssystem eingesetzt, das auf modifizierten MFCC-Merkmalen beruht.

In [1] wurde das Modell der dynamischen Adaption vorgestellt. Ein ähnliches Modell ist auch in [6] zu finden. Bei diesem Modell wird eine Ein-Ausgabe-Funktion definiert, die sich an dem Antwortverhalten der Basilarmembran [7] orientiert und als Zielvorgabe dient. Nach jeder Eingabe reguliert ein Adaptionsmechanismus einen internen Zustand, um die Ausgabe langsam an den Zielwert anzupassen. Der Zielwert entspricht der Empfindung, die ein Signal beim Hörer auslöst, sofern das Signal über einen längeren Zeitraum anliegt und nicht von einem vorherigen Signal in der Empfindung beeinflußt wird. Der interne Zustand dagegen ist vom zeitlichen Verlauf des Signals abhängig. Die Differenz zwischen dem dynamischen internen Zustand und dem statischen Zielwert ermöglicht dann die Berechnung des Maskierungseffektes.

Da das Modell der dynamischen Adaption, das mit allen notwendigen Details beschrieben wurde, bereits in einem auf MFCC-Merkmalen beruhenden Spracherkennungssystem eingesetzt wurde und dabei keine großen Änderungen in der Merkmalsberechnung notwendig waren, wird das bestehende MFCC-basierte Spracherkennungssystem mit diesem Modell erweitert. Allerdings wurden einige Änderungen in der Anwendung des Modell vorgenommen, die im folgenden Abschnitt erläutert werden.

2.2. Eingesetztes Modell

Um das Modell der dynamischen Adaption mit den in [1] angegebenen Parameterwerten anwenden zu können, müssen die Energien der einzelnen Frequenzgruppen zunächst in Dezibel umgerechnet werden. Die dazu benötigte Bezugsschallintensität ist im Allgemeinen jedoch unbekannt, da diese vom Aufnahmekanal abhängt. Deshalb wird auf den Sprachdaten dynamisch ein Energiehistogramm erstellt, mit dessen Hilfe die Gesamtenergie so normiert wird, daß das Maximum bei 60 dB liegt. Der für die Normierung berechnete additive Offset wird dann als Bezugsschallintensität verwendet.

Der Zielwert von 60 dB entspricht dem minimalen Sprachpegel [8] und wurde gewählt, um eine zu starke Maskierung zu verhindern. Nähme man stattdessen z.B. 70 dB (mittlerer Sprachpegel) als Zielwert, würden aufgenommene Sprachsignale, die tatsächlich nur bei 60 dB lagen, zu stark maskiert werden. Wird dagegen ein aufgenommener Sprachpegel von tatsächlich 70 dB auf 60 dB normiert, so findet zwar eine zu geringe Maskierung statt, eine Unterschätzung der Maskierung ist jedoch immer noch besser als gar keine Maskierung oder gar eine Überschätzung.

Nach einer Reihe von Experimenten wurden zwei Änderungen in der Anwendung des Modells vorgenommen.

Zum einen bleibt die vom Modell vorgenommene Adaption von Frequenzanregungen für die weitere Merkmalsberechnung unberücksichtigt. Durch die Adaption werden nämlich längere Frequenzanregungen, die sich z.B. durch einen Vokal oder einen Konsonanten ergeben, langsam gedämpft. Das führt dazu, daß der Unterschied zwischen verschieden langen Frequenzanregungen geringer wird und dadurch ein Teil der Sprachcharakteristik verloren geht. Außerdem wird starkes Hintergrundrauschen vom Modell adaptiert. Dies bewirkt zwar eine Dämpfung des Rauschens, gleichzeitig werden aber auch Sprachsignale gedämpft, die während des Hintergrundrauschens auftreten. Dadurch wird die Vergleichbarkeit von Sprachsignalen bei unterschiedlich starkem Hintergrundrauschen verringert.

Außerdem berücksichtigt das Modell nur Werte oberhalb von sogenannten statischen Schwellwerten, die sich an dem Bereich orientieren, in dem i.d.R. Sprachsignale auftreten. Dadurch wird jedoch der dynamische Bereich der Daten eingeschränkt. Deshalb bleiben diese Schwellwerte bei der weiteren Merkmalsberechnung ebenfalls unberücksichtigt. Allerdings muß für jede Frequenzgruppe ein Minimum, das dynamisch auf den Sprachdaten erfaßt wird, eingeführt werden, um zu verhindern, daß die Maskierung zu sehr niedrigen Werten, vor allem in den Pausen zwischen den Sprachsignalen, führt.

3. Anpassung der Merkmalsberechnung

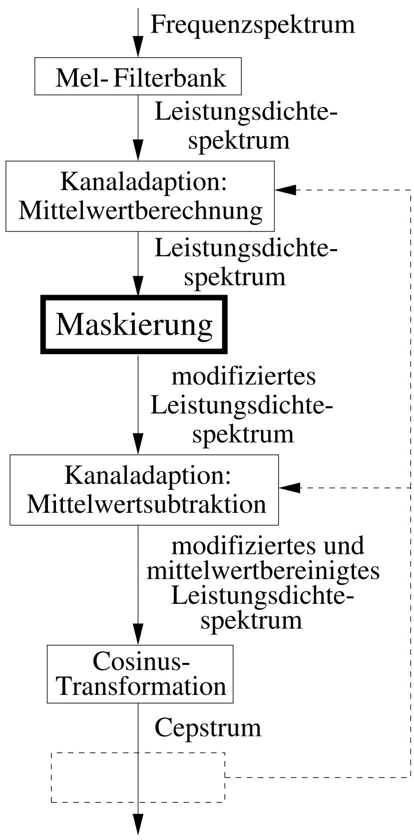

Abbildung 2: *Anpassungen in der Merkmalsberechnung zur Berücksichtigung des Effektes der Vorwärtsmaskierung*

Das beschriebene Modell läßt sich sehr einfach in einem auf MFCC-Merkmalen beruhenden Spracherkennungssystem einsetzen. Wird allerdings eine Kanaladaption auf der Basis einer sogenannten *cepstralen Mittelwertsbereinigung* [9] durchgeführt, so kommt es zu einer negativen Wechselwirkung zwischen der Kanaladaption und der Vorwärtsmaskierung. Diese Wechselwirkung läßt sich dadurch erklären, daß zum Zeitpunkt der cepstralen Mittelwertsbereinigung nicht mehr das benötigte (transformierte) Sprachsignal, sondern ein von der Maskierung modifiziertes Signal vorliegt und deshalb eine fehlerhafte Kanaladaption durchgeführt wird.

Um diese negative Wechselwirkung zu verhindern, ist eine Änderung im Ablauf der Merkmalsberechnung notwendig (siehe Abbildung 2): die Kanaladaption erfolgt nicht mehr auf der Basis des Cepstrums, sondern auf der Basis des Leistungsdichtespektrums (logarithmierte Ausgabe der Mel-Filterbank). Dazu werden die für die Kanaladaption benötigten Mittelwerte zunächst auf dem Leistungsdichtespektrum berechnet, jedoch noch nicht subtrahiert.

Nach dieser Berechnung wird das Modell für die Vorwärtsmaskierung auf das Leistungsdichtespektrum angewendet, und erst danach erfolgt die Subtraktion der Mittelwerte vom modifizierten (maskierten) Leistungsdichtespektrum. Auf diese Weise läßt sich eine negative Wechselwirkung zwischen Kanaladaption und Vorwärtsmaskierung weitestgehend verhindern.

4. Evaluierung

Das mit der Maskierung erweiterte Spracherkennungssystem wurde auf umfangreichen Sprachdaten aus verschiedenen Anwendungsdomänen evaluiert.

Die erste Anwendungsdomäne umfaßt die akustische Steuerung von nicht sicherheitsrelevanten Fahrzeugfunktionen, z.B. dem Radio oder eines Autotelefons. Bei den verwendeten SLACC-Daten (**S**poken **La**nguage **C**ar **C**ontrol) handelt es sich um Sprachdaten, die in Bielefeld in verschiedenen Fahrzeugtypen bei unterschiedlichen Geschwindigkeiten und Witterungsverhältnissen mit jeweils zwei verschiedenen Mikrofonen aufgenommen wurden: Mit einem Nahbesprechungsmikrofon (Mikro 1) sowie einem am vorderen rechten Dachpfosten befestigten Mikrofon (Mikro 2). Die Datensätze umfassen jeweils 10984 Äußerungen von insgesamt 22 Sprechern (vorgelesene Sätze), die für die Evaluierungen aufgeteilt wurden: 18 Sprecher mit 9243 Äußerungen wurden für das Training verwendet, die verbleibenden 4 Sprecher mit 1741 Äußerungen bildeten die Teststichproben. Der Wortschatz für Trainings- und Teststichproben beinhaltet jeweils insgesamt 658 Wortformen.

Daneben erfolgte eine Evaluierung auf dem deutschen *Verbmobil*- und dem englischen *Wall-Street-Journal*-Datensatz. Der Verbmobil-Datensatz besteht aus über 32 Stunden spontan gesprochenem Trainingsmaterial mit einem Wortschatz von 5336 Wortformen sowie etwa 35 Minuten Testmaterial [10]. Für Wall-Street-Journal (WSJ0) stehen ca. 15 Stunden Trainings- und etwa 40 Minuten Testmaterial (vorgelesene Sätze aus der gleichnamigen Wirtschaftszeitung) zur Verfügung, der Wortschatz[2] für die Evaluierung umfaßt dabei 4986 Wortformen [11].

Für das Spracherkennungssystem werden 39 Merkmale verwendet, die sich aus der Gesamtenergie und den ersten 12 Cepstralkoeffizienten sowie den ersten und zweiten Ableitungen zusammensetzen. Die Klassifikation erfolgt auf der Basis von semikontinuierlichen HMMs mit linearer Struktur und variabler Zustandszahl. Als Wortuntereinheiten werden Triphone ohne Generalisierung, aber mit automatischem Zustandsclustering eingesetzt. Für die Evaluierungen auf dem Verbmobil- und dem WSJ0-Datensatz werden zusätzlich Bi-Gramm-Sprachmodelle (Perplexität: 64.4 auf Verbmobil, 109.6 auf WSJ0) verwendet. Eine detailliertere Beschreibung des Spracherkennungssystems ist in [12] zu finden.

Merkmale	SLACC (Mikro 2)	SLACC (Mikro 1)	WSJ0	Verbmobil
MFCC	25,8 %	17,5 %	12,9 %	21,8 %
maskierte MFCC	23,6 %	16,5 %	12,1 %	21,6 %

Tabelle 1: *Vergleich der Wortfehlerrate des Spracherkennungssystems ohne bzw. mit Maskierung auf den unterschiedlichen Teststichproben.*

In Tabelle 1 sind die Wortfehlerraten auf den verschiedenen Stichproben angegeben. Die Ergebnisse zeigen, daß die Maskierung zu einer signifikanten Reduktion der Wortfehlerrate auf den SLACC-Datensätzen führt, wobei die Verbesserung auf den mit dem zweiten Mikrofon

[2] Die phonotypischen Transkriptionen des verwendeten Wortschatzes wurden auf der Basis des "Carnegie Mellon Pronouncing Dictionary" Version 0.6 erzeugt.

aufgenommenen Daten deutlicher ausfällt (relative Verbesserung von 8,5 %) als auf den Daten des Nahbesprechungsmikrofons (relative Verbesserung von 5,7 %). Die Verbesserung auf dem WSJ0-Datensatz ist ebenfalls noch signifikant (relative Verbesserung von 6,2 %), auf dem Verbmobil-Datensatz dagegen nicht mehr [13]. Das angewendete Modell führt also insbesondere dann zu einer Reduzierung der Wortfehlerrate, wenn die Sprachdaten vergleichsweise schlecht bzw. mit vielen Störsignalen behaftet sind.

Neben der Reduktion der Wortfehlerrate zeigt die Maskierung einen weiteren positiven Effekt: Die Varianz der Wortfehlerrate innerhalb der Teststichproben nimmt ab. Dies wird in Tabelle 2 und 3 verdeutlicht.

	Teststichprobe SLACC (Mikro 1)						
	Sprecher 1		Sprecher 2	Sprecher 3		Sprecher 4	
Merkmale	Set 1 (1286)	Set 2 (1425)	Set 1 (1065)	Set 1 (1149)	Set 2 (1207)	Set 1 (1568)	Set 2 (1767)
MFCC	28,95 %	21,08 %	25,09 %	21,32 %	23,09 %	5,26 %	6,01 %
maskierte MFCC	26,02 %	19,90 %	23,89 %	22,27 %	20,13 %	5,45 %	6,12 %

Tabelle 2: *Vergleich der Wortfehlerrate des Spracherkennungssystems ohne bzw. mit Maskierung für die Testsets der verschiedenen Sprecher. Die Angaben in Klammern geben die Anzahl der Wörter in dem jeweiligen Testset an.*

Berechnet man die Varianz der Wortfehlerrate auf den in Tabelle 2 angegebenen Daten (die Wortfehlerrate jedes Testsets wird dazu mit der jeweiligen Anzahl der Wörter gewichtet), zeigt sich, daß diese bei dem System mit Maskierung deutlich geringer ist (siehe Tabelle 3). Auch auf den anderen Teststichproben ist eine deutliche Abnahme der Varianz zu beobachten[3]. Die Maskierung führt also dazu, daß das System robuster gegenüber den Aufnahmebedingungen (Sprecher, Mikrofon, Umgebung) wird.

Merkmale	SLACC (Mikro 2)	SLACC (Mikro 1)	WSJ0	Verbmobil
MFCC	88,3	81,4	10,1	7,2
maskierte MFCC	72,8	66,2	6,2	5,6

Tabelle 3: *Vergleich der Varianz der Wortfehlerrate des Spracherkennungssystems ohne bzw. mit Maskierung auf den unterschiedlichen Teststichproben.*

5. Zusammenfassung

In diesem Beitrag wurde ein auf MFCC-Merkmalen beruhendes Spracherkennungssystem mit einem Modell erweitert, das die Berücksichtigung des Effektes der Vorwärtsmaskierung ermöglicht und das sich ohne große Änderung der Merkmalsberechnung einbinden läßt. Das erweiterte Spracherkennungssystem zeigt auf den meisten eingesetzten Sprachdaten, vor allem bei vorhandenen Störsignalen, eine signifikante Abnahme der Wortfehlerrate und auf allen Sprachdaten eine Abnahme der Varianz in der Wortfehlerrate zwischen den verschiedenen Aufnahmebedingungen.

[3] Die absolute Varianz auf Verbmobil sowie auf Wall-Street-Journal ist wesentlich kleiner, weil diese Testsets mehr Sprecher umfassen und die Differenz zwischen der jeweils besten und schlechtesten Wortfehlerrate deutlich geringer ausfällt als auf SLACC.

Literatur

[1] B. Strope, A. Alwan: *A model of dynamic auditory perception and its application to robust word recognition*, *IEEE Trans. on Speech and Audio Processing*, Bd. 5, Nr. 5, 1997, S. 451–464.

[2] C. H. Yim, O. C. Au, W. Wan, C. L. Keung, C. C. Fung: *Auditory spectrum based features (ASBF) for robust speech recognition*, in *Proc. Int. Conf. on Spoken Language Processing*, Bd. II, Beijing, Okt. 2000, S. 979–982.

[3] J. Tchorz, B. Kollmeier: *A model of auditory perception as front end for automatic speech recognition*, *J. Acoust. Soc. Am.*, Bd. 106, Nr. 4, 1999, S. 2040–2050.

[4] Y. Ito, H. Matsumoto, K. Yamamoto: *Forward Masking on a generalized logarithmic scala for robust speech recognition*, in *Proc. Int. Conf. on Spoken Language Processing*, Bd. III, Beijing, Okt. 2000, S. 530–533.

[5] T. Dau, D. Püschel, A. Kohlrausch: *A quantitative model of the „effective“ signal processing in the auditory system. I: Model structure*, *J. Acoust. Soc. Am.*, Bd. 99, Nr. 6, 1996, S. 3615–3622.

[6] Z. Liu, X. Wu, B. Zhen, H. Chi: *An auditory feature extraction method based on forward-masking and its application in robust speaker identification and speech recognition*, in *Proc. Int. Conf. on Spoken Language Processing*, Bd. II, Beijing, Okt. 2000, S. 266–269.

[7] B. Johnstone, R. Patuzzi, G. Yates: *Basilar membrane measurements and the travelling wave*, *Hearing Research*, Bd. 22, 1986, S. 147–153.

[8] E. Zwicker, H. Fastl: *Psychoacoustics*, Springer-Verlag, Berlin Heidelberg, 1999.

[9] M. Westphal: *The use of cepstral means in conversational speech recognition*, *Proc. European Conf. on Speech Communication and Technology*, Bd. 3, 1997, S. 1143–1146.

[10] K. Kohler, G. Lex, M. Pätzold, M. Scheffers, A. Simpson, W. Thon: *Handbuch zur Datenaufnahme und Transliteration in TP 14 von VERBMOBIL - 3.0*, Institut für Phonetik und digitale Sprachverarbeitung, Kiel, 1994.

[11] D. B. Paul, J. M. Baker: *The Design for the Wall Street Journal-based CSR Corpus*, in *Speech and Natural Language Workshop*, Morgan Kaufmann, 1992.

[12] G. A. Fink: *Developing HMM-based Recognizers with ESMERALDA*, in V. Matoušek, P. Mautner, J. Ocelíková, P. Sojka (Hrsg.): *Lecture Notes in Artificial Intelligence*, Bd. 1692, Springer, Berlin Heidelberg, 1999, S. 229–234.

[13] E. Paulus, M. Lehning: *Die Evaluierung von Spracherkennungssystemen in Deutschland*, *Fortschritt der Akustik - DAGA 94*, Bd. A, 1994, S. 147–156.