BESTIMMUNG DER OPTIMALEN HMM PARAMETER ZUR ROBUSTEN, PHONEMBASIERTEN SPRACHERKENNUNG

Harald Finster, Hans-Günter Hirsch

Hochschule Niederrhein
hans-guenter.hirsch@hs-niederrhein.de

Abstract: Zur Erweiterung eines bestehenden Sprachdialogsystems, das bisher auf einer wortbasierten Spracherkennung beruht, wird eine Modellierung von Wörtern als Kette von Phonem HM-Modellen (Hidden Markov) untersucht. Es wird die Abhängigkeit der Erkennungsgüte von der Anzahl der Zustände eines Modells sowie die Anzahl der Gauß-Verteilungen ermittelt. Dies wird sowohl für eine Erkennung isoliert gesprochener Kommandowörter als auch für Ziffernketten durchgeführt. Zur Extraktion der akustischen Merkmale aus dem Sprachsignal werden die beiden von ETSI standardisierten Verfahren eingesetzt. Neben einer Modellierung der Phoneme als Monophone wird auch eine Modellierung als Triphone unter Berücksichtigung der umgebenden Lautklassen betrachtet. Es lassen sich Wortfehlerraten unterhalb von 1 % für die Aufgabe des Erkennens 50 verschiedener Kommandowörter sowie Wortfehlerraten unterhalb von 3 % für eine Ziffernkettenerkennung erzielen. Bei im Auto aufgenommenen, gestörten Ziffern liegt die Fehlerrate bei etwa 7 %.

1 Einleitung

Zur Demonstration der Einsatzmöglichkeiten und der Effizienz eines Verfahrenes zur robusten Spracherkennung wurde ein Sprachdialogsystem entwickelt [1], mit dem beispielsweise telefonbasierte Informationsdienste angeboten werden können. Die Spracherkennung beschränkt sich dabei bisher auf den Einsatz von Referenzmustern, mit denen ganze Wörter modelliert werden und die auch aus ganzen Wörtern trainiert wurden. Dadurch ist die Erweiterung eines solchen Erkennungssystem zur Erkennung von neuen Wörtern, die in den üblicherweise zum Training verwendeten Sprachdatensammlungen nicht enthalten sind, mit einem relativ hohen Aufwand verbunden. Zum Training derartiger Wörter werden mindestens etwa 100 Äußerungen entsprechend vieler verschiedener männlicher und weiblicher Sprecher benötigt, um bereits eine halbwegs gute Erkennung in der ersten Betriebsphase eines neuen Dienstes zu gewährleisten. Der Aufwand zur Aufnahme und zur Aufbereitung der Sprachdaten mit einer zusätzlichen Beschreibung der Sprachsegmentgrenzen in Textform ist dabei erheblich.

Um diesen Aufwand zu reduzieren, bietet sich der Einsatz einer phonembasierten Erkennung an. Ein Ziel dieser Untersuchungen ist dabei das Training von HM-Modellen aller im Deutschen vorhandenen Phoneme. Ein weiteres Ziel ist das Ermitteln der Leistungsfähigkeit einer auf Ganzwortmodellen beruhenden Spracherkennung, wobei die Modelle durch die Aneinanderreihung der entsprechenden Phoneme erzeugt werden.

Zur Merkmalsextraktion und zur Erkennung werden im Rahmen dieser Untersuchungen die in vielen praktischen Realisierungen eingesetzte Cepstralanalyse und die Erkennung mit HM-Modellen (Hidden Markov) unter Einsatz des Viterbi Algorithmus verwendet.

Speziell werden zur Merkmalsextraktion die beiden von der ETSI Arbeitsgruppe „Aurora“ ausgewählten und standardisierten Verfahren [2],[3] eingesetzt. Das Training der HM-Modelle und die Erkennung werden mit den entsprechenden Programmen der HTK (Hidden Markov Model Toolkit) Softwaresammlung [4] durchgeführt. Im Mittelpunkt dieser Untersuchungen steht dabei das Ermitteln und die Festlegung der günstigsten Parameter zur

Modellierung von Phonemen mit HM-Modellen bezüglich der Anzahl von Zuständen eines HM-Modells sowie der Anzahl der Gauß-Verteilungen für jeden akustischen Parameter und jeden Zustand des HM-Modells. Zur Modellierung der Laute werden die beiden Ansätze verglichen, jeden Laut unabhängig von der lautlichen Umgebung als *Monophon* oder in Abhängigkeit der vorausgehenden und der nachfolgenden Lautklasse als *Triphon* zu modellieren.

Es werden zunächst der Gesamtaufbau des Erkennungssystems mit den eingesetzten Verfahren zur Merkmalsextraktion und dem Modellieren der Laute als HM-Modelle vorgestellt. Im Anschluss werden einige Details der Aufbereitung der Sprachdaten, um sie zum Training verwenden zu können, sowie des Trainings der HM-Modelle selbst dargestellt. Nach einer Vorstellung der Erkennungsexperimente werden die erzielten Ergebnisse präsentiert und ausgewertet.

2 Aufbau des Erkennungssystems

Ein Überblick über den im Rahmen dieser Untersuchungen eingesetzten Gesamtaufbau zur automatischen Spracherkennung wird in Abbildung 1 gegeben.

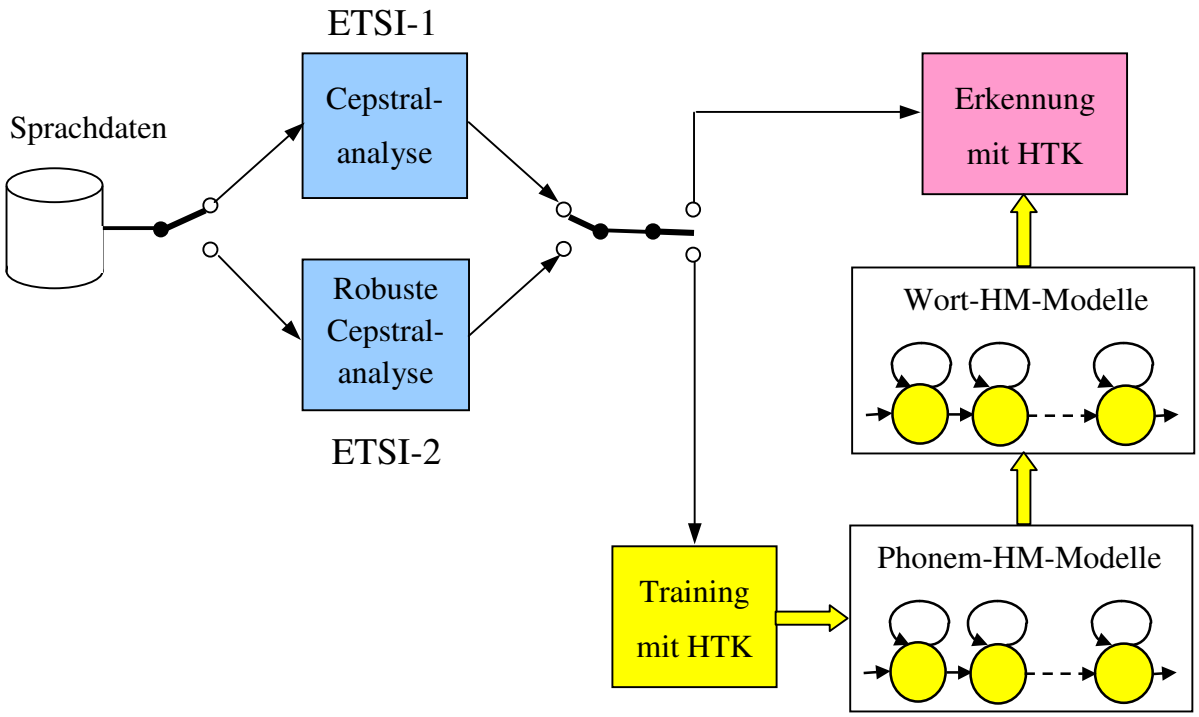

Abbildung 1 - Gesamtaufbau

2.1 Merkmalsextraktion

Zur Extraktion der zur Erkennung benutzten akustischen Merkmale wird eine der beiden bei ETSI standardisierten Vorgehensweisen eingesetzt.

Bei dem auch in der zeitlichen Abfolge als erstem standardisiertem Verfahren [2] handelt es sich um eine auf einer 23 kanaligen MEL Filterbank beruhenden Cepstralanalyse ohne zusätzliche Verarbeitungsschritte, die über die üblicherweise verwendete Vorgehensweise zur Generierung von MEL-Frequenz-Cepstral-Koeffizienten (MFCC) hinausgehen. Im zeitlichen Abstand von 10 ms werden 13 Cepstral-Koeffizienten einschließlich des 0. Koeffizienten bestimmt, wobei der 0. Koeffizient als ein die Kurzzeit-Energie charakterisierender Parameter angesehen werden kann. Zusätzlich wird noch ein weiterer Kurzzeit-Energiewert aus den

quadrierten Abtastwerten des jeweils analysierten Signalabschnitts bestimmt. Im Rahmen dieser Untersuchungen werden 12 Cepstral-Koeffizienten ohne Berücksichtigung des 0. Koeffizienten und der aus den Abtastwerten bestimmte Energie-Koeffizient verwendet. Diese 13 akustischen Parameter werden um die sogenannten Delta und Delta-Delta Koeffizienten ergänzt, die als 1. bzw. 2. Ableitung des zeitlichen Verlaufs jedes der 13 Parameter anzusehen sind. Die Bestimmung der Delta und Delta-Delta Koeffizienten wird gemäß der auch bei HTK verwendeten Vorgehensweise durchgeführt. Damit bestehen die Ausgangswerte dieser Merkmalsextraktion, auf die im weiteren mit dem Kürzel ETSI-1 Bezug genommen wird, aus 100 Vektoren je Sekunde, wobei jeder Vektor 39 Komponenten beinhaltet.

Das zweite von ETSI standardisierte Verfahren [3] stellt eine Erweiterung des ersten Verfahrens um einige weitere Verarbeitungsblöcke dar. Die zusätzlichen Verarbeitungsschritte wurden mit dem Ziel der Gewährleistung einer robusteren Erkennung eingeführt. Mit dem standardisierten Verfahren wurden bei einem Vergleich mehrerer Vorschläge, die von der ETSI Arbeitsgruppe mit dem Namen “Aurora“ evaluiert wurden, die besten Erkennungsergebnisse für verschiedene Erkennungsaufgaben erzielt [5]. Die zusätzliche Verarbeitung beinhaltet eine auf einer Wiener Filterung beruhende Reduktion stationärer Hintergrundstörungen sowie ein Verfahren zur blinden Schätzung eines unbekannten Übertragungs-Frequenzgangs. Letztlich werden auch bei diesem Verfahren 12 Cepstral-Koeffizienten (ohne den 0.) und 1 weiterer, die Kurzzeit-Energie beschreibender Parameter bestimmt. Die Bestimmung der Delta und Delta-Delta Koeffizienten erfolgt auf eine geringfügig unterschiedliche Weise [3]. Auch bei diesem Verfahren, das durch den Kürzel ETSI-2 beschrieben wird, werden somit 100 Vektoren je Sekunde mit 39 Komponenten je Vektor erzeugt.

Beide Programme, mit denen die zuvor beschriebenen Formen der Signalanalyse realisiert wurden, speichern die akustischen Merkmale in dem bei HTK festgelegten Format ab.

2.2 Modellierung der Phoneme

In der Trainingsphase werden mit Hilfe der entsprechenden HTK Programme die HM-Modelle für die in Tabelle 1 aufgelisteten 45 Phoneme, die in einer Sampa ähnlichen Notation beschrieben werden, sowie zwei Modelle “sil“ und “sp“ zur Beschreibung langer und kurzer Sprachpausen bestimmt.

Phoneme	a: e: i: o: u: a E I O U u y: ae oe Oe On E: ar an @ h f v z x s S C Z m n l r j Y N t g p d k b aI OY aU

Tabelle 1 – Liste der verwendeten Phoneme

Dabei werden die Anzahl der Zustände je Modell und die Anzahl der Gauß-Verteilungen zur Beschreibung jedes akustischen Parameters in einem Zustand variiert. Konkret wird die Anzahl der Zustände zwischen 3 und 9 (3, 5, 7 und 9) und die Anzahl der Gauß-Verteilungen zwischen 1 und 16 (1, 2, 4, 8 und 16) variiert. Bei HM-Modellen mit 3 Zuständen ist das Überspringen eines Zustand nicht erlaubt, wohingegen bei Modellen mit mehr als 3 Zuständen das Überspringen eines Zustand erlaubt ist.

Neben der Modellierung der Phoneme als Monophone, bei denen keine Abhängigkeit von den umgebenden Lauten berücksichtigt wird, wird auch eine Modellierung als Triphone untersucht. Dabei wird die Modellierung allerdings nicht in der Form vorgenommen, dass man ein Modell eines Lauts von dem speziellem vorhergehendem und dem nachfolgendem Laut abhängig macht, sondern nur von der Lautklasse des vorhergehenden und des nachfolgenden Lauts. Dabei wird die in Tabelle 2 angegebene grobe Einteilung aller Laute in 5 Klassen berücksichtigt.

Kürzel	Klasse	Phoneme
V	Vokale	a: e: i: o: u: a E I O U u y: ae oe ar an E: @
F	Frikative	h f v z x s S C
N	Nasale	m n l r j Y Z N On
S	Plosive	t g p d k b
D	Umlaute	aI OY aU

Tabelle 2 – Einteilung in Lautklassen

Diese Vorgehensweise zur Triphon Modellierung besitzt den Vorteil, dass man letztlich die Gesamtzahl der Modelle auf etwa 700 beschränken kann und dass man bei den zum Training verwendeten Sprachdaten nicht auf das Problem einer unzureichenden Anzahl von Äußerungen eines Triphons trifft.

Aus den Phonemmodellen werden durch einfache Aneinanderreihung HM-Modelle zur Beschreibung ganzer Wörter generiert. Dazu wird die lautsprachliche Beschreibung, wie sie in einem Aussprache-Lexikon vorhanden ist, benötigt. Es wurde ein Programm geschrieben, das die im HTK Format abgespeicherten Modelle einliest, die Aneinanderreihung gemäß eines Eintrags im Aussprache-Lexikon vornimmt und das resultierende Wortmodell wieder im HTK Format abspeichert.

Die so generierten Wortmodelle werden dann in der eigentlichen Erkennungsphase zur Einzelwort- oder zur Wortkettenerkennung herangezogen, wobei die Erkennung mit dem entsprechenden HTK Programm vorgenommen wird.

3 Training

Zum Training der Phonemmodelle wird ein Teil der über Telefon aufgenommenen deutschen SpeechDat Sprachdatensammlung verwendet. Es werden Telefondaten zum Training benutzt, da letztlich auch eine telefonbasierte Anwendung des Sprachdialogsystems angestrebt wird. Diese Aufnahmen weisen die im Bereich der Telefonie und bei einer Sprachübertragung im Festnetz üblichen Störungen auf, die im wesentlichen aus eventuell vorhandenen Hintergrundgeräuschen beim Anrufer bestehen. Insgesamt wurden 500 männliche und 500 weibliche Sprecher aufgezeichnet.

Etwa 32000 der in der SpeechDat Sammlung vorhandenen Aufnahmen werden zum Training verwendet. Dies entspricht einer Gesamtlänge der Aufnahmen von etwa 48 Stunden. Die Aufnahmen beinhalten beispielsweise einzelne Kommandowörter, Wortketten wie beispielsweise Ziffernfolgen bis hin zu ganzen Sätzen. Ein Problem bestand in der fehlenden Beschreibung der zeitlichen Lautgrenzen innerhalb jeder Aufnahme.

Die lautsprachliche Beschreibung kann relativ leicht mit einem Aussprache-Lexikon aus der wortbasierten Beschreibung jeder Aufnahme erstellt werden. Die Festlegung der Phonemgrenzen in einer Äußerung ist hingegen deutlich aufwendiger und schwieriger. Dies wurde im Rahmen dieser Untersuchungen im wesentlichen mit einer „erzwungenen“ Erkennung realisiert. Bei einer „erzwungenen“ Erkennung wird dem Spracherkenner mitgeteilt, was in einer Äußerung gesprochen wurde, so dass auch nur die entsprechende Folge von HM-Modellen erkannt werden kann. Dabei wird auch eine zeitliche Zuordnung der Sprachabschnitte zu den in der Aufnahme enthaltenen Wörtern oder Lauten erzeugt.

Es stellte sich jedoch heraus, dass eine mehrstufige Vorgehensweise notwendig ist, um verwendbare Ergebnisse zu erzielen. Für etwa 400 Aufnahmen, die das Buchstabieren von Wörtern beinhalten, wurden von den Autoren mit Hilfe eigener Programme die Lautgrenzen

manuell ermittelt. Dabei beinhalten diese Aufnahmen nur eine aus 25 Phonemen bestehende Untermenge aller zu trainierenden Phoneme. Mit Hilfe dieser Lautgrenzen konnten in einer ersten Phase initiale HM-Modelle für die in den Aufnahmen enthaltenen Phoneme erzeugt werden.

Damit wurden dann in einer weiteren Trainingsphase unter Verwendung einer deutlich größeren Anzahl von Sprachaufnahmen, die schon die Artikulation ganzer Wörter beinhalten, eine größere Anzahl von Phonemen als HM-Modelle trainiert. Mit den so trainierten Modellen wurde dann eine „erzwungene“ Erkennung durchgeführt, die eine zeitliche Beschreibung der Wortgrenzen innerhalb einer Aufnahme liefert. Als wichtiger Aspekt stellte sich auch das Vorhandensein eines verlässlichen Modells für die Sprachpausen heraus, um damit die Sprachgrenzen zuverlässig festlegen zu können.

Durch mehrfaches iteratives Wiederholen der zuvor beschriebenen Vorgehensweise mit einem erneuten Training der Phonemmodelle und einer nachfolgenden „erzwungenen“ Erkennung konnte letztlich für das gesamte, zum Training verwendete Sprachdatenmaterial eine recht verlässliche Beschreibung der zeitlichen Lautgrenzen erzeugt werden.

Unter Verwendung dieser Lautgrenzeninformation wird mit Hilfe der zum Training zur Verfügung stehenden HTK Programme ein Training der Monophone durchgeführt. Aus der vorhandenen Beschreibung der Lautgrenzen in dem dafür vorgesehenen HTK Textformat kann auch eine entsprechende Beschreibung für die Triphone gewonnen werden, so dass damit ein Training der Triphone durchgeführt werden kann.

4 Experimente zur Spracherkennung

Um die Verwendbarkeit der trainierten Phonemmodelle und die Leistungsfähigkeit einer darauf basierenden Spracherkennung festzustellen, wurden einige Erkennungsexperimente definiert.

Bei dem ersten Experiment handelt es sich um eine Einzelworterkennung von 52 deutschen Kommandowörtern (z.B. “Hilfe“, “Stop“, “ja“, “nein“, …). Es werden Sprachdaten zur Erkennung verwendet, die zu der zum Training benutzten Sprachdatensammlung gehören und auch zum Training benutzt wurden. Insgesamt werden etwa 7650 Aufnahmen der Kommandowörter betrachtet. Diese relativ einfache Erkennungsaufgabe spiegelt eine vermutlich bei Einsatz des Sprachdialogsystems häufig anzutreffende Aufgabenstellung wieder. Im folgenden wird das Kürzel **CMD** als Kurzbezeichnung für dieses Experiment verwendet.

Das zweite Experiment besteht in einer Erkennung deutscher Ziffernketten. Dabei handelt es sich um eine deutlich schwierigere Erkennungsaufgabe, für die in der Regel wortbasierte HM-Modelle eingesetzt werden. Neben Sprachdaten, die aus der zum Training verwendeten Datenbasis stammen, werden in diesem Fall auch Daten untersucht, die aus anderen Datenbasen stammen und in einer anderen akustischen Umgebung aufgenommen wurden. Damit kann die Leistungsfähigkeit der phonembasierten Erkennung bezüglich der Robustheit in anderen akustischen Umgebungen festgestellt werden. Insgesamt werden 3 verschiedene Datensätze benutzt.

Der erste Datensatz, der im folgenden durch das Kürzel **S1** beschrieben wird, besteht aus etwa 2100 Aufnahmen der Trainingsdatenbasis. In den 2100 Aufnahmen sind insgesamt etwa 10000 gesprochene Ziffern enthalten.

Die Sprachdaten des zweiten Datensatzes wurden nicht über Telefon aufgenommen, sondern bei Nahbesprechung eines Mikrofons in einer relativ ruhigen Umgebung. Somit unterscheidet sich die akustischen Umgebung deutlich von der der Trainingsdaten. Dieser zweite Datensatz **S2** umfasst Aufnahmen von etwa 90 Sprechern, bei denen teilweise starke dialektale

Färbungen vorhanden sind. Es handelt sich um etwa 1900 Aufnahmen mit insgesamt 7000 gesprochenen Ziffern.

Bei dem dritten Datensatz, auf den im folgenden durch das Kürzel **S3** Bezug genommen wird, handelt es sich um den Teil der deutschen SpeechDatCar Datensammlung, die auch bereits zur Evaluierung in der ETSI Arbeitsgruppe eingesetzt wurden. Diese Sprachdaten wurden in der störschallerfüllten Umgebung eines Kraftfahrzeugs aufgenommen. Damit sind diese Daten insbesondere dazu geeignet, die Leistungsfähigkeit bei Vorhandensein von Hintergrundstörungen zu untersuchen. Es handelt sich um etwa 3000 Aufnahmen mit insgesamt 16500 gesprochenen Ziffern.

Einen Überblick über die zur Erkennung verwendeten Sprachdaten wird in Abbildung 2 gegeben.

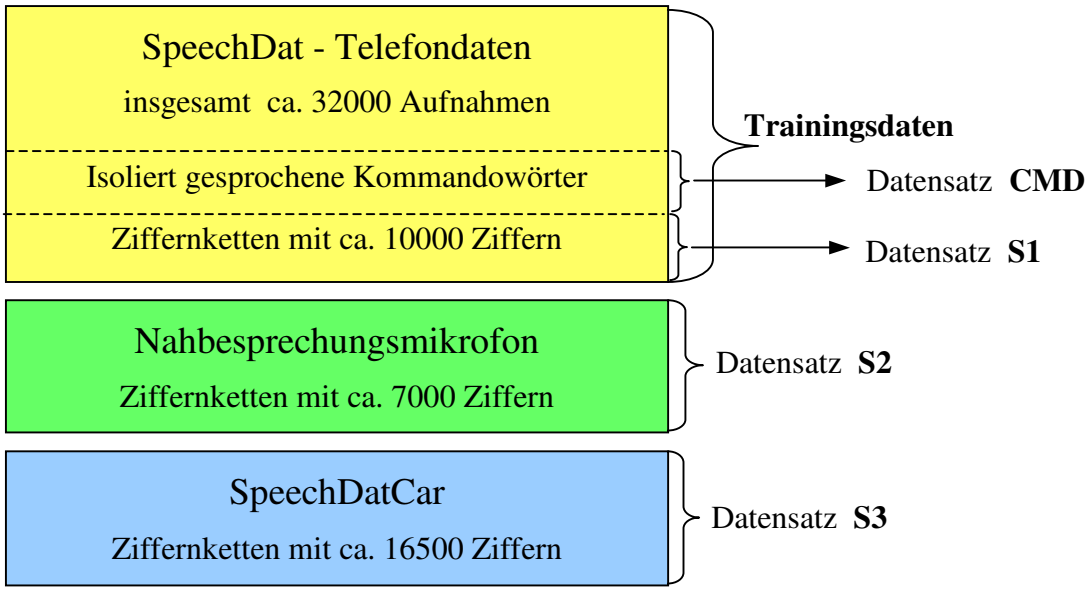

Abbildung 2 – Überblick der zum Training und zur Erkennung eingesetzten Sprachdaten

5 Spracherkennungergebnisse

Zunächst werden die Ergebnisse bei Verwendung von Monophon Modellen präsentiert.

In den Tabellen 3 und 4 sind die Wortfehlerraten für die Erkennung der isoliert gesprochenen Kommandowörter des Datensatzes CMD zusammengestellt. Dabei wird die Variation der Anzahl der Zustände eines HM-Modells im Bereich von 3 bis 7 und der Anzahl der Gauß-Verteilungen im Bereich von 1 bis 16 für jeden akustischen Parameter in einem Zustand betrachtet. Diese Fehlerraten beinhalten auch die Fehler, die durch zusätzlich erkannte Modelle (Einfügungen) bzw. durch nicht erkannte Wörter (Auslöschungen) zustande gekommen sind. Die Ergebnisse in Tabelle 3 werden bei Einsatz der Merkmalsextraktion ETSI-1, die Ergebnisse in Tabelle 4 bei Einsatz der Merkmalsextraktion ETSI-2 erzielt. Es zeigt sich zunächst prinzipiell, dass die Fehlerraten für eine größer werdende Anzahl von Zuständen eines HM-Modells sowie eine zunehmende Anzahl von Gauß-Verteilungen in der Regel geringer werden. Diese Ergebnisse wurden tendenziell auch so erwartet.

Die Fehlerraten sind für diesen Fall einer Einzelworterkennung bei Vergleich der Merkmalsextraktion ETSI-2 mit der Merkmalsextraktion ETSI-1 nur geringfügig besser. Dies ist auch darauf zurückzuführen, dass die zur Erkennung verwendeten Daten aus der gleichen akustischen Umgebung wie die Trainingsdaten stammen. In einem solchen Anwendungsfall macht sich die erhöhte Robustheit des ETSI-2 Verfahrens nicht sonderlich bemerkbar.

Zahl der HMM Zustände	Zahl der Gauß-Verteilungen je Zustand				
	1	2	4	8	16
3	10,4	5,4	4,6	3,5	2,9
5	8,1	5,5	4,2	3,0	2,7
7	5,5	3,5	2,4	1,9	1,7

Tabelle 3 – Wortfehlerraten (%) für die Einzelworterkennung (CMD) mit der ETSI-1 Merkmalsextraktion

Zahl der HMM Zustände	Zahl der Gauß-Verteilungen je Zustand				
	1	2	4	8	16
3	9,9	4,9	4,0	2,8	2,7
5	7,1	4,3	3,5	2,8	2,4
7	4,8	3,1	2,2	1,5	1,3

Tabelle 4 – Wortfehlerraten (%) für die Einzelworterkennung (CMD) mit der ETSI-2 Merkmalsextraktion

Ähnliche Ergebnisse stellen sich für die Erkennung der Ziffernketten des Datensatzes S1 ein, was die Abhängigkeit von der Anzahl der Zustände eines HM-Modells, die Anzahl der Gauß-Verteilungen als auch den Vergleich der beiden Merkmalsextraktionsverfahren ETSI-1 und ETSI-2 betrifft. Die Fehlerraten nehmen nur grundsätzlich größere Werte an, was auf die wesentlich schwierigere Aufgabe einer Wortketten-Erkennung zurückzuführen ist.

Ein Vergleich der mit den beiden Verfahren zur Merkmalsextraktion erzielten Ergebnisse kann an Hand der Bilder in Abbildung 3 angestellt werden. Es werden die erzielten Fehlerraten bei Variation der Anzahl der Modellzustände für eine Anzahl von 16 Gauß-Verteilungen dargestellt. Dabei ist zunächst die unterschiedliche Skalierung der Ordinaten zu beachten. Die Fehlerraten für die Datensätze S2 und S3 sind bei dem ETSI-2 Verfahren mit Werten im Bereich von 5 bis 15 % deutlich niedriger im Vergleich zu ETSI-1 mit Werten im Bereich von 20 bis 40%. Bei diesen Datensätzen, die Daten aus anderen akustischen Umgebungen beinhalten, zeigt sich der Gewinn, der durch das Extrahieren robuster Merkmale mit dem ETSI-2 Verfahren erzielt werden kann. Besonders deutlich wird der Gewinn bei Betrachtung der Kurven für den Datensatz S2, für den mit ETSI-2 größtenteils sogar bessere Ergebnisse als für den Datensatz S1 erzielt werden können, obwohl die Sprachdaten von S1 auch zum Training verwendet wurden.

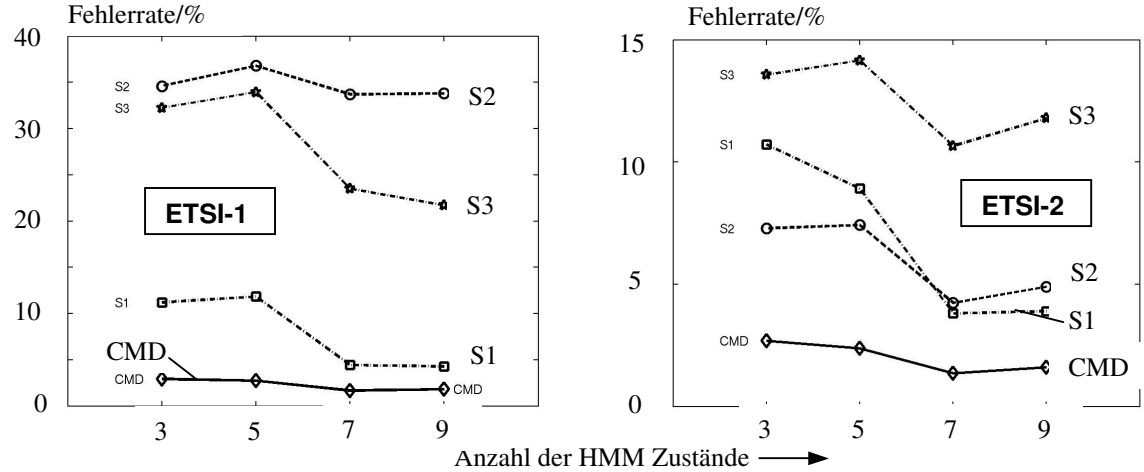

Abbildung 3 – Wortfehlerraten (%) für alle Datensätze als Vergleich zwischen ETSI-1 und ETSI-2

Es zeigt sich, dass die niedrigsten Fehlerraten bei Modellen mit 7 Zuständen auftreten. Bei einer weiteren Erhöhung der Modellzustände ergibt sich nahezu in allen Fällen eine

Verschlechterung. Mit einer Anzahl von 9 Zuständen ist offensichtlich keine gute Modellierung kurzer Phoneme mehr möglich.

Abschließend werden in Abbildung 4 die Fehlerraten für den Vergleich einer Modellierung mit Triphonen im Vergleich zur Modellierung mit Monophonen verglichen. Dabei werden die Ergebnisse für eine feste Anzahl von 7 Modellzuständen bei einer Variation der Anzahl der Gauß-Verteilungen präsentiert, wobei die Merkmale mit dem ETSI-2 Verfahren extrahiert wurden. Es wird deutlich, dass durch die Modellierung der Phoneme als Triphone unter Berücksichtigung der umgebenden Laute in allen Fällen eine bessere Erkennung erzielt werden kann.

Der Abbildung lässt sich auch entnehmen, dass die Erhöhung der Anzahl der Gauß-Verteilungen von 8 auf 16 keinen größeren Gewinn bringt. Resümierend lässt sich festhalten, dass durch eine HM-Modellierung mit 7 Zuständen je Modell und mit 8 Gauß-Verteilungen eine gute Erkennung gewährleistet werden kann..

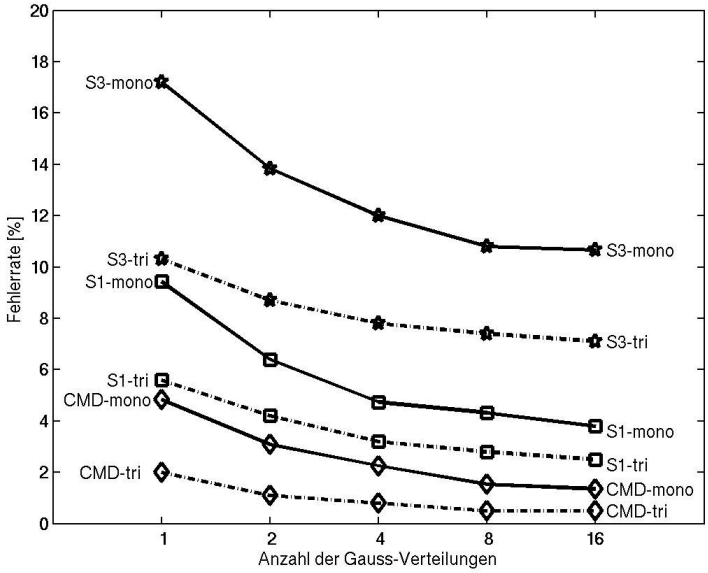

Abbildung 4 – Wortfehlerraten (%) für Monophon ("mono") und Triphon ("tri") Modelle mit 7 Zuständen bei Anwendung der ETSI-2 Merkmalsextraktion

Literatur

[1] Hirsch, H.G.: Realisierung eines Sprachdialogsystems mit einer robusten Sprach-erkennung. 15. Konferenz zur elektronischen Sprachsignalverarbeitung, Cottbus, 2004

[2] ETSI standard document: Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Extended front-end feature extraction algorithm; Compression algorithm; Back-end speech reconstruction algorithm, ETSI ES 202 211 v1.1.1 (2003-11), Nov. 2003

[3] ETSI standard document: Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithm; Back-end speech reconstruction algorithm, ETSI ES 202 212 v1.1.1 (2003-11), Nov. 2003

[4] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK book – version 2.2, Entropic, 1999

[5] Pearce, D., Mauray, L., Noe, B. et al.: Evaluation of a noise robust DSR front-end on Aurora databases. 7th International Conference on Spoken Language Processing, Denver, 2002, pp. 17-20