EIN SPRECHERADAPTIVER HIDDEN-MARKOV ERKENNER FÜR GROßE WORTSCHÄTZE

Karl Trottler, Telefunken Systemtechnik GmbH, 7900 Ulm
Fritz Class, Peter Regel, Daimler Benz Forschungszentrum, 7900 Ulm

In diesem Beitrag wird ein sprecheradaptives Erkennungssystem beschrieben, das sich in einer kurzen Lernphase automatisch an einen neuen und unbekannten Sprecher adaptieren kann. Die Sprecheradaption basiert auf einer Transformationsvorschrift für die Merkmalsvektoren, die mit Hilfe eines minimalen Fehlerquadrat-Ansatzes optimiert wird. Das Verfahren kann sowohl in Spracherkennungssystemen auf der Basis der dynamischen Zeitnormierung als auch in Hidden-Markov Erkennern eingesetzt werden.

1. Einleitung

Schnelle Sprecheradaptionsverfahren gewinnen zunehmend an Bedeutung in Spracherkennungssystemen, die ein großes Vokabular verarbeiten. Der bislang übliche Weg, das System an einen neuen Benutzer zu adaptieren, geht davon aus, daß jedes Wort des Vokabulars mindestens einmal in einer Trainingsphase eingesprochen werden muß. Dies ist jedoch keine praktikable Lösung, wenn - wie bei Erkennungssytemen basierend auf Hidden-Markov Modellen oder neuronalen Netzen - große Datenmengen für das Training der Klassifikationsparameter zur Verfügung gestellt werden müssen. Es sind daher neue Methoden erforderlich, um ein Spracherkennungssystem für große Wortschätze an einen neuen Sprecher zu adaptieren.
In diesem Beitrag wird ein Verfahren beschrieben, bei dem jeder Merkmalsvektor mit Hilfe einer spektralen Abbildung in einen anderen Merkmalsvektor transformiert wird. Aufgabe einer Optimierung ist es, eine geeignete Transformation in einer möglichst kurzen Adaptionsphase zu ermitteln. Der hier verwendete Ansatz beruht auf dem Kriterium des minimalen mittleren Fehlerquadrates [1] und resultiert in mathematisch angebbaren Beziehungen, die in einer Systemrealisierung mit Signalprozessoren quantitativ auswertbar sind.

Im folgenden wird das Adaptionsverfahren, das auf einer zweiseitigen spektralen Abbildung von Merkmalsvektoren beruht, kurz beschrieben. Es werden die Experimente und deren Ergebnisse dargestellt. Abschließend werden einige Aspekte zur Realisierung eines sprecheradaptiven HMM-Erkenners, der auf diesem Adaptionsverfahren basiert, vorgestellt.

2. Das Adaptionsverfahren

2.1 Transformation in einen gemeinsamen Merkmalsraum

Die Idee des Verfahrens, das auf Überlegungen von Grenier et al. [2] zurückgeht, beruht darauf, die Merkmalsvektoren $\underline{X}$ bzw. $\underline{Y}$ eines neuen Sprechers bzw. eines Referenzsprechers, der den Spracherkenner vorab trainiert hat, in einen gemeinsamen Merkmalsraum zu transformieren. Ausgehend von dem Ansatz des minimalen mittleren quadratischen Abstandes

$$D = E[\ (\mathbf{P}_L \cdot \underline{X} - \mathbf{P}_R \cdot \underline{Y})^T \cdot (\mathbf{P}_L \cdot \underline{X} - \mathbf{P}_R \cdot \underline{Y})\ ] \qquad (1)$$

der transformierten Merkmalsvektoren $\underline{x} = \mathbf{P}_L \cdot \underline{X}$ und $\underline{y} = \mathbf{P}_R \cdot \underline{Y}$, besteht nun die Aufgabe darin, die beiden Transformationsmatrizen $\mathbf{P}_L$ und $\mathbf{P}_R$ zu bestimmen. Um die triviale, aber nicht befriedigende Lösung $\mathbf{P}_L = \mathbf{P}_R = \underline{0}$ von vornherein auszuschließen, führen wir die Nebenbedingung

$$E[x_k^2] = E[y_k^2] = 1 \qquad (2)$$

ein, d.h. wir fordern, daß die Komponenten x_k und y_k der beiden transformierten Vektoren x und y gleiche Varianz besitzen. Unter Berücksichtigung

dieser Normierungsbedingung kann das Variationsproblem komponentenweise formuliert werden entsprechend der Beziehung

$$D_k = E[(x_k - y_k)^2] = 2(1 - E[x_k^* y_k]) \stackrel{!}{=} \min. \qquad (3)$$

Der Fehlerterm D_k ist somit minimal, wenn die Komponenten x_k und y_k der transformierten Vektoren maximal korreliert sind. Die Lösung dieses Problemes resultiert aus der sog. "canonical correlation analysis" [3,4]. Das Ergebnis ist ein verallgemeinertes Eigenwertproblem, das mit Hilfe der singular value decomposition einer Matrix numerisch auswertbar ist. Diese Matrix enthält die Auto- bzw. Kreuzkovarianzmatrizen der beiden Sprecher, die in einer kurzen Trainingsphase anhand korrespondierender Merkmalsvektoren abgeschätzt werden. Der neue Sprecher spricht zu diesem Zweck die gleichen Worte wie der Referenzsprecher nach.

2.2 Nichtlinear erweiterte Merkmalsvektoren

Aufgrund von nichtlinearen Abhängigkeiten zwischen den einzelnen Merkmalskomponenten ist eine Adaption an linear transformierten Vektoren nur begrenzt leistungsfähig. Andererseits läßt sich eine lineare Merkmalstransformation in weitgehend geschlossener Form angeben. Man kann nun beide Überlegungen kombinieren, indem man eine lineare Transformation auf quadratisch erweiterte Merkmalsvektoren $\underline{v}_Q = (v_1, v_2, \ldots, v_K, v_1^2, v_1 v_2, \ldots, v_K^2)$, d.h. auf Polynomvektoren zweiten Grades anwendet. Bei der Berechnung der optimalen Transformationsmatrizen wird dabei - wie oben geschildert - verfahren.

3. Testbedingungen

Die Sprachsignale wurden tiefpaßgefiltert und mit 12 kHz abgetastet. Anschließend erfolgte die Berechnung von mel-cepstral Koeffizienten (MCC) in einem zeitlichen Raster von 10 msec, wobei für die weitere Verarbeitung K=10 MCCs pro Merkmalsvektor verwendet wurden. Das Testvokabular bestand aus 100 gebräuchlichen deutschen Worten, die von vier männlichen und einem weiblichen Sprecher vorgesprochen worden waren. Es wurden von jedem Sprecher zwei Stichproben mit je 100 Worten an zwei verschiedenen Tagen aufgezeichnet. Für das Training des HMM-Erkenners lag darüber hinaus eine Stichprobe von ca. 1000 Worten (15 Minuten) eines Sprechers vor, der demzufolge als Referenzsprecher definiert wurde. Die übrigen vier Sprecher galten als neue, an das Erkennungssystem zu adaptierende Sprecher.

4. Experimente und Ergebnisse

Für den Test der Adaptionsverfahren wurden zwei prinzipiell unterschiedliche Klassifikationsverfahren eingesetzt: ein abstandsmessender Erkenner auf der Basis der dynamischen Zeitnormierung (DTW) und ein Hidden-Markov Erkenner (HMM). Das HMM-System ist im Detail in [5] beschrieben. Die einzelnen Worte werden repräsentiert durch eine Verkettung von Hidden-Markov Modellen von Wortuntereinheiten. Die phonetischen Wortgraphen, deren Knoten die Wortuntereinheiten darstellen, werden mit Hilfe eines Regelwerkes automatisch aus der orthographischen Beschreibung erzeugt. Die HMMs werden durch diskrete Emissionswahrscheinlichkeiten beschrieben. Das zur Diskretisierung erforderliche sprecherabhängige Codebuch setzt sich aus 128 Codebuchvektoren zusammen. Zusätzlich wird die Energie mit Hilfe von drei Symbolen quantisiert.

Abb. 1 zeigt die Adaptionsergebnisse für beide Erkennungsverfahren. Die Qualität der Sprecheradaption wird dabei bewertet anhand eines Vergleiches der sprecherabhängigen Fehlerrate (SA) mit den sprecheradaptiven Fehlerraten für lineare (GRE) bzw. quadratisch erweiterte (GRE_Q) Merkmalsvektoren sowie mit der Fehlerrate ohne Adaption (OA). Es ist offensichtlich, daß die Fehlerraten aufgrund der Sprecheradaption drastisch verringert werden können. Das Verfahren GRE_Q liefert in beiden Fällen bessere Ergebnisse als das lineare

GRE-Verfahren. Die sprecherabhängige Fehlerrate läßt sich durch die Sprecheradaption bis auf eine geringe Abweichung von etwa 2% erreichen. Vergleichen mit der Fehlerrate ohne Adaption verringert sich die Fehlerrate im sprecheradaptiven Fall für den DTW-Erkenner um den Faktor 6 bzw. für den HMM-Erkenner um den Faktor 3.

Ein weiteres Ziel der Sprecheradaption ist es, den Trainingsaufwand möglichst gering zu halten. Die Anzahl der einzusprechenden Worte in der Adaptionsphase ist somit ein weiteres Kriterium zur Abschätzung der Leistungsfähigkeit des untersuchten Adaptionsverfahrens. Es wurde daher die Anzahl der zur Adaption herangezogenen Worte variiert. In Abb. 2 sind die resultierenden Fehlerraten für beide Adaptionsverfahren im Vergleich zur Fehlerrate ohne Adaption und zur sprecherabhängigen Fehlerrate für den DTW-Erkenner dargestellt. Es wird offensichtlich daß bei dem GRE_Q-Verfahren bereits ca. 40 Worte für die Adaption des neuen Sprechers ausreichen.

5. Realisierung eines sprecheradaptiven HMM-Erkenners

Die Algorithmen für den sprecheradaptiven HMM-Erkenner wurden auf einem PC-basierten Demonstrationssystem implementiert. Ein Blockschaltbild ist in Abb. 3 dargestellt.
Die Erfassung der Sprachdaten (Tiefpaßfilterung, A/D-Wandlung mit 16 Bit bei 12 kHz Abtastfrequenz) und deren Vorverarbeitung (Bildung von Kanalvektoren, Sprach-/Pause-Segmentierung) ist auf einer käuflichen PC-Einschubkarte basierend auf dem Festkomma-Signalprozessor TMS320C25 realisiert.
Eine zweite Karte basiert auf dem Gleitkomma-Signalprozessor DSP32C von AT&T und ist zuständig für die Durchführung der Adaption und zur HMM-Klassifikation. Zunächst werden die Cepstralkoeffizienten gebildet, die dann Normierungsoperationen unterzogen werden. In der Adaptionsphase werden anhand dieser normierten Koeffizienten die beiden Transformationsmatrizen $\mathbf{P}_L$ und $\mathbf{P}_R$ gebildet. Vor der Erkennphase werden dann die auf den Referenzsprecher angepaßten Codebuchvektoren mit Hilfe der Matrix $\mathbf{P}_R$ transformiert. Die normierten Cepstralvektoren des neuen Sprechers werden mit der Matrix $\mathbf{P}_L$ transformiert, mit dem transformierten Codebuch diskretisiert und an den HMM-Modul übergeben, der die Klassifikation durchführt. Die daraus resultierenden N besten Wortkandidaten werden zur weiteren Verarbeitung an den PC-Steuerprozessor transferiert. Das System ist konzipiert für die Erkennung eines Vokabulars von ca. 1000 Worten.
Weitere Arbeiten konzentrieren sich auf die Realisierung einer mitlaufenden Sprecheradaption, auf die Realisierung eines sprecherunabhängigen HMM-Erkenners auf der Basis von Wortuntereinheiten sowie auf die Erweiterung für eine kontinuierliche Spracherkennung.

6. Literatur

[1] F.Class,A.Kaltenmeier,P.Regel,K.Trottler: Fast speaker adaptation for speech recognition systems. ICASSP 1990, Albuquerque, New Mexico, USA.
[2] K.Choukri,G.Chollet,Y.Grenier: Spectral transformation through canonical correlation analysis for speaker adaptation in ASR. ICASSP 1986, Tokio.
[3] T.W.Anderson: An introduction to multivariate Statistical Analysis. J.Wiley & Sons, New York, 1958.
[4] F.Class,P.Regel,K.Trottler: Speaker adaptation for recognition systems with a large vocabulary. Proc. of MELECON 1989, Lissabon.
[5] F.Class et al.: Speaker adaptive word verification using hidden Markov models of sound units for a recognition system with a large vocabulary. Proc. of 7th FASE Symposium SPEECH 88, Edinburgh, pp. 23-30.

Diese Arbeit wurde teilweise durch das Bundesministerium für Forschung und Technologie gefördert (ITM 8801). Allein die Autoren sind für den Inhalt verantwortlich.

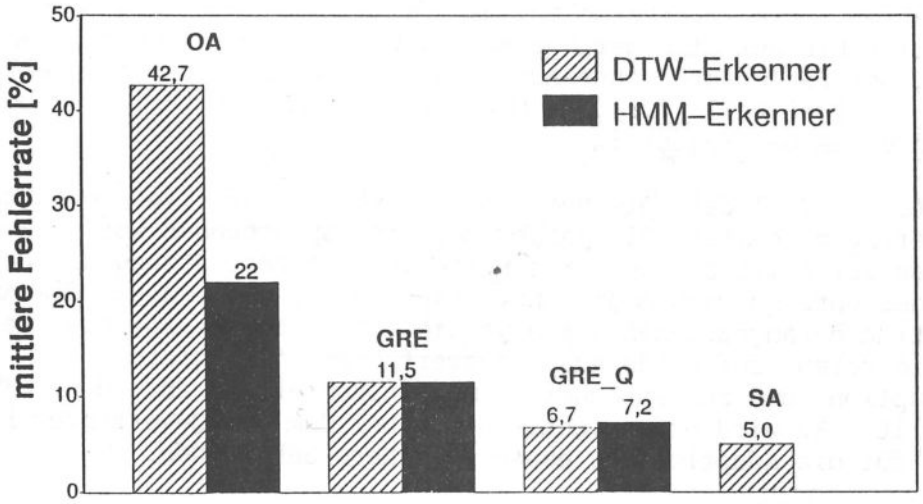

Abb. 1: Mittlere Fehlerraten mit und ohne Adaption, sprecherabhängige Fehlerrate, 100 Worte in der Trainingsphase (ca. 1.5 min Sprache)

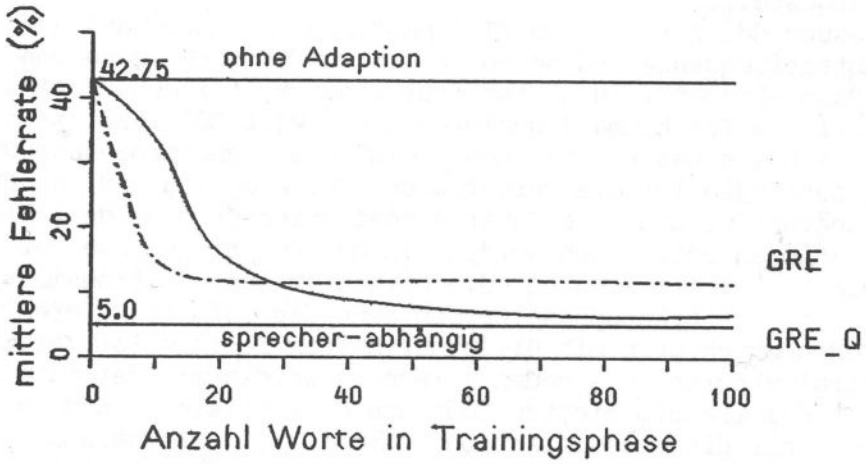

Abb. 2: Mittlere Fehlerrate in Abhängigkeit von der Anzahl der Trainingsworte, DTW-Erkenner

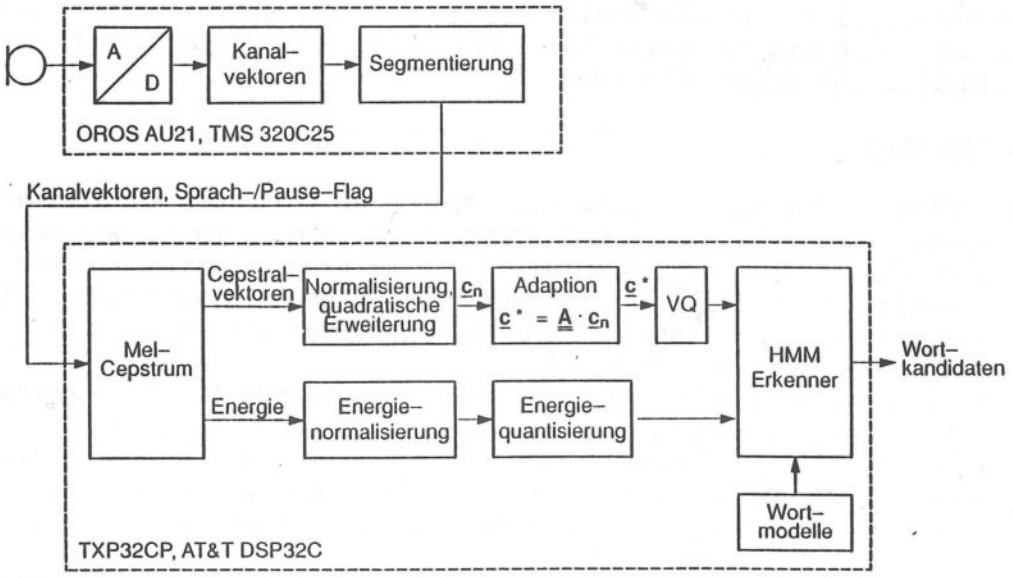

Abb. 3: Blockschaltbild des realisierten sprecheradaptiven HMM-Erkenners für große Wortschätze