BMBF

Verbmobil

Verbundvorhaben

Automatische Wortsegmentierung mit semikontinuierlichen Hidden Markov Modellen

Michael Lehning, Rainer Grünheid

TU Braunschweig

Vm

Report 71
Juli 1995

Juli 1995

Michael Lehning, Rainer Grünheid

Institut für Nachrichtentechnik
Technische Universität Braunschweig
Schleinitzstr. 22
38092 Braunschweig

Tel.: (0531) 391 - 2453
Fax: (0531) 391 - 8218
e-mail: m.lehning@tu-bs.de

Gehört zum Antragsabschnitt: 14.3 Werkzeuge zur Segmentation und Etikettierung

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 N 0 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Automatische Wortsegmentierung mit semikontinuierlichen Hidden Markov Modellen

Michael Lehning, Rainer Grünheid
Institut für Nachrichtentechnik, TU Braunschweig

1 Einleitung

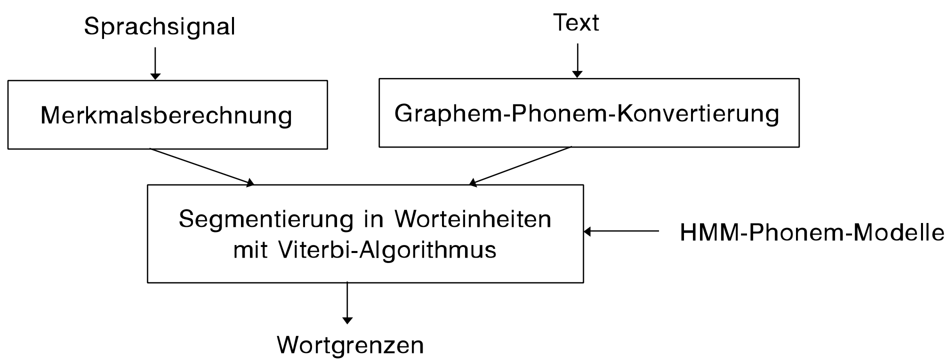
Die manuelle Wortsegmentierung mittels auditiver und visueller Kontrolle ist eine zeit- und personalintensive Arbeit. Aus diesem Grund wird nach Verfahren gesucht, die diesen Vorgang zumindest teilweise automatisieren.
Bei dem hier beschriebenen Ansatz wird ein Spracherkennungssystem auf der Basis von semikontinuierlichen Hidden Markov Modellen (SCHMM) [HUANG1990] zur automatischen Segmentierung von Äußerungen in Wort- bzw. Wortuntereinheiten auf Signalebene benutzt. Zu diesem Zweck wird aus der orthographischen Transliteration mittels phonologischer Regeln oder einem Aussprachelexikon die phonetische Umschrift des Satzes in kanonischer Form abgeleitet. Aus der so bestimmten phonetischen Umschrift wird durch Verkettung der zugeordneten SCHMM-Phonemmodelle ein Gesamtmodell für den zu segmentierenden Satz erstellt. Mittels des Viterbi-Algorithmus wird nun die bestmögliche Abbildung zwischen den Modellzuständen und den Merkmalsvektoren des Satzes berechnet (erzwungene Erkennnung). Aus dem Ergebnis der Abbildung werden dann die Wortgrenzen ermittelt.

2 Systemüberblick

Das Segmentierungssystem besteht aus folgenden Komponenten (siehe Bild 1):

1. Merkmalsextraktion und -transformation
2. Ermittlung der phonetischen Umschrift des Satzes aus der Rechtschrift
3. Bestimmung der Modellreihenfolge aus der phonetischen Umschrift
4. Ausrichtung der verketteten Phonemmodellzustände mittels Viterbi-Algorithmus bei vorgegebener Modellreihenfolge
5. Ermittlung der Wortgrenzen aus der Ausrichtung

Bild 1: Systemübersicht



3 Graphem-Phonem-Konvertierung

Augenblicklich wird die phonetische Umschrift einem Aussprachelexikon entnommen, in dem jedem Rechtschriftwort die entsprechende phonetische Umschrift zugeordnet ist. Ein Rechtschriftwort wird somit in eine Folge von Phonemen umgesetzt.

4 Merkmalsberechnung

Die das Sprachsignal beschreibenden Merkmale werden durch eine Kurzzeitsignalanalyse des Sprachsignals gewonnen. Die Analyse erfolgt im 10ms Raster. Die Analysefenster besitzen eine Breite von 16ms. Bei einer Abtastfrequenz von 16 KHz liegen somit 256 Signalwerte innerhalb eines Fensters. Diese Werte werden durch Multiplikation mit einem 256 Werte breitem Hamming-Fenster gewichtet. Für jedes Analysefenster werden folgende Merkmale berechnet:

- 12 Cepstralparameter, die näherungsweise aus 16 LPC-Koeffizienten berechnet werden
- 12 zeitliche Ableitungen der Cepstralparameter
- Log. Energie der gewichteten Abtastwerte im Fenster
- Ableitung der logarithmierten Energie

5 Codebuchgenerierung

Bei den semikontinuierlichen HMM-Modellen wird die Auftrittswahrscheinlichkeit fr einen Vektor $\vec{x}_t$ durch eine multimodale Gaußverteilung Φ, die sich aus der gewichteten Einzelsumme von L Gaußfunktionen mit den Verteilungsparametern $\vec{v}_0 \ldots \vec{v}_{L-1}$ ergibt, geschätzt:

$$f(\vec{x}_t|\Phi) = \sum_{l=0}^{L-1} b_l f(\vec{x}_t|\vec{v}_l)$$

$\vec{v}_l$ beschreibt dabei die Verteilungsparameter der Gaußverteilung l, d.h. Mittelwertvektor μ_l und die Kovarianzmatrix Σ_l, von der nur die Hauptdiagonalelemente berücksichtigt werden. Die Werte fr b_l werden während der Adaptionsphase aus den Initialisierungswerten nachgeschätzt. Die Verteilungsparameter μ_l und Σ_l sind für alle Zustände gleich. Die Verteilungsparameter werden aus dem Ergebnis einer Vektorquantisierung der Trainingsvektoren in L Cluster geschätzt. Die Clusterung erfolgt nach dem LBG-Algorithmus [GRAY1984]. Die Zentroiden des Clusters l werden mit dem Mittelwertvektor μ_l gleichgesetzt. Die Varianzen des Clusters l werden mit den Hauptdiagonalemente von Σ_l gleichgesetzt. Die Codebuchparameter können während der Modelladaption nachgeschätzt werden, was allerdings den Zeitaufwand beim Training erhöht.

6 Erkennungseinheiten

Als Erkennungseinheiten werden Phonemmodelle mit drei Zustnden (engl. States) verwendet, bei denen Übergänge in den nächsten und in den übernächsten Zustand möglich sind.

7 Modellinitialisierung

Zur Schätzung der initialen Übergangswahrscheinlichkeiten zwischen den M Zuständen (hier $M = 3$) innerhalb eines SCHMM-Phonem-Modells wird jedes handsegmentierte Phonem, das durch P Merkmalsvektoren $\vec{x}_0 \ldots \vec{x}_{P-1}$ beschrieben wird, automatisch in M zusammenhängende Bereiche $S_0 \ldots S_{M-1}$ segmentiert. Jedes Segment S_i umfaßt die Merkmalsvektoren $\vec{x}_{s_i} \ldots \vec{x}_{s_{i+1}-1}$ (d.h. s_i ist der Index für die linke Grenze des Segmentes S_i und $s_0 = 0$, $s_M = P$). Der Zentroidvektor $\vec{C}_i$ für das Segment S_i berechnet sich aus dem arithmetischen Mittel der dem Segment S_i zugeordneten Vektoren:

$$\vec{C}_i = \frac{1}{s_{i+1} - s_i} \sum_{k=s_i}^{s_{i+1}-1} \vec{x}_k$$

Die ”Kosten” für die Zusammenfassung der Vektoren $\vec{x}_{s_i} \ldots \vec{x}_{s_{i+1}-1}$ ergeben sich aus der Summe der quadratischen Distanzen zum Zentroiden $\vec{C}_i$ gewichtet mit der Länge des Segmentes:

$$D_i = \frac{1}{s_{i+1} - s_i} \sum_{k=s_i}^{s_{i+1}-1} (\vec{x}_k - \vec{C}_i)^2$$

Das Segmentierungsverfahren beruht auf dem Prinzip der Dynamischen Programmierung [PARSONS1986].
Als Optimierungskriterium D wird die minimale Gesamtdistanz über alle Einzelkosten D_i gewählt:

$$D = \sum_{i=0}^{M-1} D_i \qquad \longrightarrow \qquad \text{min}$$

Beginnend mit S_0 werden die Kosten für alle möglichen Endpunkte von S_0 berechnet. Im nächsten Schritt werden alle möglichen Endpunkte von S_1 berechnet und dabei für jeden Endpunkt des Segmentes S_1 die im Sinne der Kostenfunktion bestmögliche Aufteilung der Vektoren $\vec{x}_{s_0} \equiv \vec{x}_0$ bis $\vec{x}_{s_2-1}$ in zwei Cluster S_0 und S_1 gesucht, wobei trivialerweise $s_2 \geq s_1$ gilt. Dieses Vorgehen wird nun rekursiv für alle weiteren Segmente S_m mit $m = 2 \ldots M-1$ durchgeführt.
Aus der so gewonnenen Aufteilung der handsegmentierten Phonembereiche werden die Selbstübergangswahrscheinlichkeiten a_{ii} und Emissionswahrscheinlichkeiten $b_i(\vec{v}_l)$ geschätzt (werden mehrere Realisierungen eines Phonems beobachtet, so wird über die bestimmten Parameter a_{ii} bzw. $b_i(\vec{v}_l)$ gemittelt).

$$a_{ii} = \frac{s_{i+1} - s_i - 1}{s_{i+1} - s_i}$$

Sind Übergänge in die nachfolgenden d States (State $i+1$ bis State $i+d$) zugelassen, so werden die Übergangswahrscheinlichkeiten aus einer geometrischen Reihe wie folgt geschätzt:

$$a_{ij} = \kappa * \delta^{(i+d-j)}(1 - a_{ii})$$

κ wird dabei so gewählt, daß gilt:

$$\kappa * \sum_{k=1}^{d} \delta^{d-k} = 1$$

Für das vorliegende Modell wurde $\delta = 9$ und $d = 2$ gewählt. κ bestimmt sich somit zu 0.1, d.h. es sind Übergänge in den nächsten und in den übernächsten Zustand möglich (double skip model) mit $a_{i(i+1)} = 0.9*(1-a_{ii})$ und $a_{i(i+2)} = 0.1*(1-a_{ii})$ [RABINER1982],[PLANNERER1992].
Die initialen Emissionswahrscheinlichkeiten für den Zustand i werden wie folgt geschätzt:

$$b_i(\vec{v}_l) = \frac{1}{s_{i+1} - s_i} \frac{\sum_{t=s_i}^{s_{i+1}-1} f(\vec{x}_t | \vec{v}_l)}{\sum_{t=s_i}^{s_{i+1}-1} \sum_{k=1}^{L} f(\vec{x}_t | \vec{v}_k)}$$

$\vec{v}_l$: Codebuchparameter für die Verteilung l, $\vec{x}_{t_i} \ldots \vec{x}_{t_{i+1}-1}$: Merkmalsvektoren, die dem Zustand (State) i zugeordnet wurden

8 Modelltraining- und test

8.1 Trainingsdaten

Das System wurde mit 3000 phonetisch balancierten Sätzen mit dem Forward-Backward-Algorithmus [HUANG1990] trainiert. Für das Training wurden keine Aussprachevarianten berücksichtigt.

8.2 Testdaten

Als Testdaten standen 100 Sätze aus der Domäne ”Intercity-Auskunft” zur Verfügung. Die automatisch bestimmten Wortgrenzen wurden manuell verifiziert und ggf. korrigiert.

8.3 Auswertung

Die Ergebnisse der automatischen Wortsegmentierung wurden qualitativ bewertet, indem folgende Fehlerkategorien eingeführt wurden:

- Phonemfehler (d.h. die Wortgrenze ist um ein Phonem in das Vorgängerwort oder in das Nachfolgewort verschoben worden)
- Silben- und Wortfehler (d.h. die gefundene Grenze liegt im Bereich einer Silbe oder eines Wortes von der tatsächlichen Wortgrenze entfernt).

Die 100 Testsätze umfaßten 989 Wörter, deren Wortgrenzen zu 80,3 % korrekt bestimmt wurden. In 19,2 % der Fälle ergab sich eine Verschiebung der Grenzen um maximal ein Phonem. Die Fehlerrate für Verschiebungen um eine Silbe oder ein Wort lag unter einem Prozent.

Literatur

[GRAY1984] Robert M. Gray: *Vector Quantization*, IEEE Acoustic, Speech and Signal Processing Magazine, April 1984, 4 — 28

[HUANG1990] X.D. Huang, Y. Ariki, M.A. Jack: *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990, ISBN 0 7486 0162 7

[LEE1989] Kai-Fu Lee: *Automatic Speech Recognition — The Development of the SPINX System.* Kluwer Academic Publishers, ISBN 0-89838-296-3

[PARSONS1986] Thomas Parsons: *Voice and Speech Processing*, McGraw-Hill Company, ISBN 0-07-048541-0

[PLANNERER1992] B. Plannerer, G. Ruske: *Recognition of Demisyllable based Units using Semicontinuous Hidden Markov Modes.* IEEE International Conference on Acoustics, Speech and Signal Processing, San Francisco(USA), 1992, p. 581 — 584

[RABINER1982] L.R. Rabiner, S.E. Levinson, M.M. Sondhi:*On the Application of Vector Quantization and Hidden Markov Models to Speaker-Independent, Isolated Word Recognition.* The Bell System Technical Journal, April 1983, 1075 — 1105

Das diesem Beitrag zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01IV101N0 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor