Optimierung eines silbenbasierten Spracherkenners

M. Tress[1], O. Schreiner[2] und G. Palm[1]

[1] *Universität Ulm, Abteilung Neuroinformatik, Oberer Eselsberg, 89081 Ulm, Deutschland*

[2] *DaimlerChrysler Forschung und Technologie/Universität Göttingen, Lonestr. 11, 89081 Ulm, Deutschland*

Einleitung

Moderne Spracherkennungsanwendungen erfordern Vokabulare mit teilweise deutlich mehr als 10.000 Wörtern. Dies ist nach wie vor eine große Herausforderung für Speicher und Laufzeit, aber auch für die Erkennungspräzision, insbesondere bei natürlichsprachlichen Systemen, bei denen unbekannte Wörter zu erwarten sind. In dieser Arbeit wird ein Ansatz der Spracherkennung vorgestellt, der auf phonetisch definierten Silben basiert. Das Wortlexikon wird hierbei durch ein Silbenlexikon ersetzt, das mit einer endlichen Anzahl Silben einen nahezu beliebigen Umfang an Wörtern abdecken kann. Der erste Teil der Arbeit beschreibt die Optimierung des Silbeninventars. Dabei wird sowohl die Ähnlichkeit von Silben in Ihren Merkmalsemissionen der HMM Zustände berücksichtigt, als auch ihre statistische Verwechselbarkeit im Erkennungsergebnis. Das Silbenlexikon wird so um jene reduziert, die phonetisch schwer unterscheidbar sind und damit keine Information für den Erkennungsprozess liefern. Der zweite Teil behandelt die Abbildung der erkannten Silbenfolge auf Wörter. Die hierfür experimentell untersuchten Methoden umfassen Levenshtein-Abstand, Assoziativspeicher und kombinierte Ansätze.

Grundlagen

Um ein Städtelexikon mit etwa 68.000 Einträgen abzudecken wird in dieser Arbeit ein Silbenlexikon mit 8116 verschiedenen Silben eingesetzt. Die Silben im Lexikon werden phonetisch getrennt, um eine möglichst geringe gegenseitige Abhängigkeit der Silben zu erhalten. Die Trennung, wie im Duden praktiziert, ist hierfür nicht geeignet. Das Ergebnis der ersten Phase ist eine Folge oder ein Hypothesengraph von Silben. Diese erkannten Silben werden in der zweiten Phase mit dem Abstandsmaß nach Levenshtein und Assoziativspeichern auf Einträge aus dem Lexikon abgebildet, um das gesprochene Wort zu ermitteln.

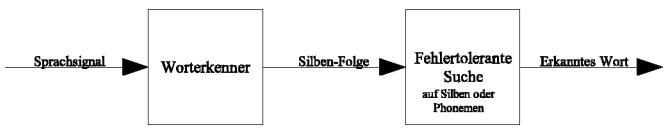

Abbildung 1: Aufbau des zweistufigen Spracherkenners

Implementierung

Für die Vektorquantisierung werden MFCC Vektoren eingesetzt, wobei in der Erkennung etwas mehr als 1600 trainierte Hidden-Markov-Modelle eingesetzt werden; es kommen Monophone, Biphone mit Links- oder Rechtskontext und Triphone mit Links- und Rechtskontext vor [1]. Zuerst wird mit der Viterbi-Dekodierung ein Jumbograph erzeugt, der in der zweiten Phase reduziert [2] und auf die wahrscheinlichste Wortfolge oder auf einen Hypothesengraphen einer bestimmten Hypothesendichte reduziert wird. Dieser Erkenner wird eingesetzt, um die wahrscheinlichste Silbenfolge zu finden, die anschließend auf einen Lexikoneintrag abgebildet wird. Hierzu werden fehlertolerante Suchmethoden benötigt, da falsch erkannte Silben enthalten sein könnten und durch die Optimierung der Silben Information verloren geht. Hierfür werden der Levenshtein-Abstand[3] und Assoziativspeicher eingesetzt[4].

Optimierung

Die Vielzahl ähnlicher Silben bewirkt eine große Verwechselbarkeit, was eine sehr niedrige Erkennungsrate zur Folge hat(Silben 4%, Wörter 5%). Ein weiterer Faktor der niedrigen Erkennung ist, dass durch das generische Silbenmodell auch unsinnige Folgen von Silben erkannt werden können. Die Erkennung kann verbessert werden, indem das Lexikon um ununterscheidbare Silben reduziert wird. Eine andere Methode ist, toleranter zu bewerten und z.B. /ha:/ als korrekt zu werten, wenn eigentlich /ha/ erwartet wird. Diese Methode kommt in der zweiten Phase zum Einsatz. Um einander ähnliche Silben zu finden, wurde die Ähnlichkeit der Emissionswahrscheinlichkeiten, die den Hidden-Markov-Modellen zugrunde liegen, eingesetzt. Die Emissionsvektoren der aufeinanderfolgenden Zustände eines Modells werden benutzt, um das Eingangssignal eines idealen Phonems zu simulieren und aus der Modellabfolge einer Silbe so das ideale Eingangssignal einer Silbe simuliert. Jede dieser idealen Silben wird anschließend als simuliertes Eingangssignal des Spracherkenners genutzt, wobei das eingesetzte Erkennungs-Lexikon alle Silben enthält. Das Ergebnis dieser Erkennung ist eine N-Best-Liste der erkannten Silben mit zugehörigen Erkennungs-Scores. Der Erkennungs-Score dient dabei als Maß der Ähnlichkeit zweier Silben, nämlich derer, die als Eingangssignal benutzt wurde und der Lexikon-Silbe, mit der sie auf diese Weise verglichen wurde. Silben-Paarungen, die auf diese Weise eine bestimmte Score-Schwelle unterschreiten, werden dann als ununterscheidbar eingestuft und eine der

beiden kann aus dem Silben-Inventar entfernt werden.

Anschließend wird mit der Verwechslungsstatistik experimentell ermittelt, welche dieser Silben der beste Repräsentant ist, d.h. welche Silbe als einzige dieser Gruppe im Silbenlexikon bleibt. Es konnte gezeigt werden, dass Silben unterschiedlich gut als Repräsentanten geeignet sind, daher ist die Suche nach der richtigen Silbe sehr wichtig für eine gute Erkennungsrate. Jede dieser ähnlichen Silben wird jeweils einmal in einem Erkennungslauf als Repräsentant eingesetzt, mit den anderen Läufen verglichen und jene Silbe als Repräsentant ausgewählt, die die beste Erkennungsrate erzielte. Für die Eingabe aus 27.220 isoliert gesprochenen Städtenamen konnte die Erkennungsrate der Silben von 3,6% auf 14%, die Erkennungsrate der Wörter von 5,2% auf 8% verbessert werden, wie in Abbildung 2 dargestellt: Entlang der X-Achse verlaufen die unterschiedlichen Optimierungs-Stufen, entlang der Y-Achse die prozentualen Erkennungsraten. Außerdem nimmt die Anzahl der Einfügungen deutlich ab, was aufgrund der ursprünglich hohen Zahl an Einfügungen ein wichtiges Kriterium für eine gute Erkennungsrate ist. In diesen Experimenten konnte der absolute Erkennungs-Score (s. Abb.2) als Schwelle bessere Ergebnisse erzielen als die Differenz der Erkennungs-Scores.

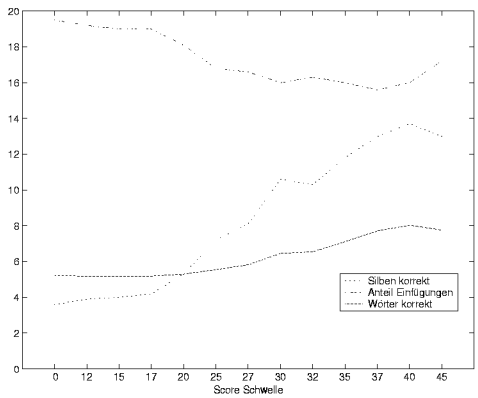

Abbildung 2: Erkennungsraten für unterschiedliche Schwellen der Optimierung

Abbildung auf Wörter

Die optimierten und erkannten Silben werden anschließend auf Einträge aus dem Lexikon abgebildet. Hierfür werden die Silben in die enthaltenen Phoneme zerlegt, da Silben für diese Aufgabe zu große Einheiten sind. Die Phonem-Erkennungsrate (42,8% Phoneme, 5,4% Wörter) liegt deutlich über der Silben-Erkennung, da eine erkannte Silbe durch ein falsches Phonem bereits als falsch gewertet wird.

Die anfängliche Erkennung liegt für den Levenshtein-Abstand bei 41% mit Eingabe der besten Silbenfolge und bei 46% mit Eingabe des Hypothesengraphen (Graph-Dichte 5). Der Assoziativspeicher mit 25,3% und 31% entsprechend. Beachtet werden muss hierbei auch die Länge der Antwortliste, da die Einträge mit größtem Score nicht weiter unterschieden werden können. Der Ganzworterkenner liegt bei einer Antwortliste der Länge 1 bei 54%. Die besten Ergebnisse konnten erzielt werden, indem das Training des Assoziativspeichers erweitert wurde und nicht nur das Lexikon, sondern Erkennungshypothesen eingesetzt wurden(vgl. Tabelle 1). Eine tolerantere Abbildung, die wahrscheinliche Phonem-Verwechslungen berücksichtigt, hat keine Verbesserungen gebracht.

Methode	Erkennung	Länge Antwortliste
Anfängliches Silben-Inventar:		
Assoz.speich.	25,3%	23
Levenshtein	41%	7
Assoz.speich.(Hyp.graph)	31%	12
Levenshtein (Hyp.graph)	46%	6
Optimiertes Silben-Inventar (absoluter Score 27):		
Assoz.speich.	27,9%	17
Assoz.speich.(Hyp.graph)	33,2%	10
Optimierung durch erweitertes Training:		
Assoz.speich.	54,5%	3
Assoz.speich.(Hyp.graph)	66,6%	1

Tabelle 1: Erkennungsraten nach Abbildung der Silben auf Wörter

Zusammenfassung

In dieser Arbeit wurde ein phonetisches Silbenlexikon eingesetzt mit dem Ziel, ein großes Wort-Lexikon zu erkennen und die Erkennungsraten zu verbessern. Hierzu wurde das Silbenlexikon durch Reduktion um kaum unterscheidbare Varianten optimiert (von 4% auf 14% Silben-Erkennung), indem mit den Emissionswahrscheinlichkeiten der HMM ähnliche Silben gesucht wurden. Die erkannten Silbenfolgen wurden in die enthaltenen Phoneme zerlegt und mit Assoziativspeichern (67%) und dem Levenshtein-Abstand (46%) auf Lexikon-Einträge abgebildet, womit eine Verbesserung im Vergleich zum Ganzworterkenner (54%) erzielt werden konnte.

Diese Arbeit wurde teilweise unterstützt durch das Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Projektes SmartWeb, Förderung 01 IMD01 D. Die Verantwortung für den Inhalt liegt bei den Autoren.

Literatur

[1] Kaltenmeier, A.: Modellbasierte Worterkennung in Spracherkennungssystemen für großen Wortschatz, Dissertation, Daimler-Benz Forschungsinstitut, 1991.

[2] Kuhn, T.; Fetter, P.; Kaltenmeier, A.; Regel-Brietzmann, P.: DP-based wordgraph pruning, in Proc. ICASSP96, 1996.

[3] Levenshtein, V.: Binary codes capable of correcting deletions, insertions and reversals - Soviet Physics Doklady, 1966.

[4] Palm, G.: On Associative Memories, Max-Planck-Institut für Biologische Kybernetik, Tübingen, 1980.