Verteilte Spracherkennung mit hybriden akustischen Modellen

Jan Stadermann und Gerhard Rigoll
Institut für Mensch-Maschine Kommunikation, Technische Universität München, Deutschland
Email: {stadermann, rigoll}@mmk.ei.tum.de

Übersicht

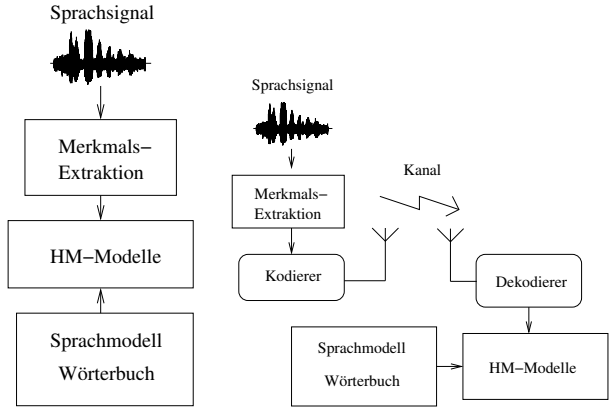
Verteilte Spracherkennung ist ein neuer Ansatz, um rechen- und speicherintensive Spracherkennertechnologie für kleine, mobile Endgeräte (Mobiltelefon, PDA) zu realisieren. Der Hauptvorteil hierbei ist die Aufspaltung der Spracherkennungsaufgabe in Merkmalsextraktion und Dekodierung. Gerade in diesem Szenario treten unerwünschte Hintergrundgeräusche auf, die in der Vorverarbeitung auf dem Client berücksichtigt werden müssen. Vorgestellt werden unterschiedliche Merkmale für die verteilte Spracherkennung sowie unterschiedliche akustische Modellansätze, welche die Übertragung der Merkmale zwischen Client und Server optimieren. Experimente mit hybriden akustischen Modellen im Umfeld der verteilten Spracherkennung lassen einige Vorteile dieser Architektur gegenüber den bekannten Gaußmodellen erkennen. Die Evaluation der unterschiedlichen Ansätze hat auf den standardisierten Tests der AURORA2 Datenbank stattgefunden.

Einleitung

Die Kommunikation mit automatischen Dialogsystemen über mobile Endgeräte gewinnt immer mehr an Bedeutung. Das Problem, das sich in diesem Szenario stellt, ist, wo das Spracherkennungssystem implementiert wird: Im *Client* stehen üblicherweise nur begrenzte Hardware-Ressourcen zur Verfügung, allerdings kann hier auf das Audiosignal ohne Übertragungsverluste zugegriffen werden. Der *Server* bietet ausreichende Rechen- und Speicherkapazität für einen komplexen Spracherkenner mit großem Vokabular und umfangreichem Sprachmodell, andererseits muß das Audiosignal dann über einen bandbegrenzten Kanal übertragen werden. Als Ausweg bietet sich hier die verteilte Spracherkennung an, bei der ein Teil des akustischen Modells auf dem Client die Audiodaten vorverarbeitet und dem übrigen Teil des Erkenners nur Merkmale über den Kanal zukommen läßt, bei denen eine große Datenreduktion ohne großen Informationsverlust stattgefunden hat. In Abbildung 1 ist ein einteiliger Spracherkenner der verteilten Architektur gegenübergestellt.

Akustische Modelle mit Gauß-HMM

Akustische Hidden-Markov Modelle (HMM) mit einer Ausgabewahrscheinlichkeitsdichte, die sich als Summe von Gaußfunktionen darstellen läßt, sind seit langem etabliert und sind auch in der verteilten Spracherkennung einsetzbar. Hier werden dafür auf dem Client mel-Cepstrum Koeffizienten (MFCC) berechnet. Der Kodierer besteht aus 7 Vektorquantisierern, die den Merkmalsvektor, wie in Bild 2 gezeigt, komprimieren [1]. Über den Kanal werden dann nur die 7 Kodebuch-Indizes übertragen, auf der Serverseite entsteht daraus dann der mit Quantisierungsfehlern behaftete Merkmalsvektor $\vec{f}$. Anschließend werden dem Merkmalsvektor noch die erste und zweite Zeitableitung hinzugefügt, was eine Verstärkung des Quantisierungsfehlers in den Ableitungskoeffizienten zur Folge hat. In den HMM werden die Merkmale dann pro Zustand mit 3 oder 6 Gaußfunktionen modelliert (s. *Experimente*), die Parameterschätzung erfolgt mit dem Baum-Welch Algorithmus.



(a) Einteiliger Erkenner (b) Verteilter Erkenner

Abbildung 1: Spracherkenner-Architekturen

Akustische Modelle mit hybriden HMM

Ein alternatives akustisches Modell benutzt ein neuronales Netz (NN) zur Schätzung von Symbolauftrittswahrscheinlichkeiten [2] mit dem Vorteil, daß das NN diskriminativ trainiert werden kann. Eine weitere Verbesserung, insbesondere für die verteilte Spracherkennung erhält man, indem alle Netzausgänge jedem HMM-Zustand i zur Verfügung stehen [3]. Die Verknüpfung ge-

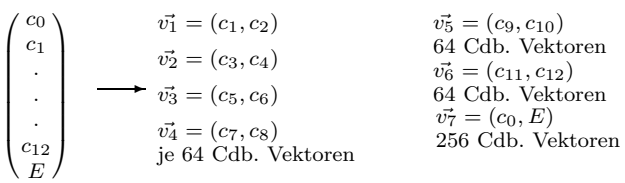


Abbildung 2: Vektorquantisierung

schieht hierbei über einen Gewichtungskoeffizienten:

$$p(\vec{x}|i) \propto \sum_{j=1}^{J} c_{ij} \cdot \frac{Pr(j|\vec{x})}{Pr(j)} \quad (1)$$

wobei $Pr(j|\vec{x})$ vom NN geliefert und $Pr(j)$ aus den Trainingsdaten ermittelt wird. Um die Datenwerte wieder auf 4,4k bit/s zu begrenzen, werden zum Einen die Wahrscheinlichkeitswerte mit dem Quantisierer nach Abb. 3 quantisiert, zum Anderen werden nur die N höchsten Werte übertragen [3]. Bei Ganzwortmodellen sind insgesamt 48 Klassen unterschieden worden (inklusive 2 Pausenklassen) [3], bei Phonemmodellen sind es 47 Klassen (45 Phoneme und 2 Pausen).

Experimente

Grundlage aller Experimente ist die AURORA2 Datenbasis, die englischsprachige Ziffernfolgen mit künstlich hinzuaddierten Geräuschen enthält. In [4] sind 11 akustische Ganzwortmodelle mit 3 Gaußdichten pro Zustand (18 Zustände pro Wort, Pausenmodell: 3 Zustände, 6 Gaußdichten) als Basissystem beschrieben. Alle Systeme werden mit verrauschten Daten trainiert, die Testdaten unterscheiden zwischen bekannten (Test A) und unbekannten (Test B) Geräuschen bei bekanntem und unbekanntem (Test C) akustischen Kanal. Es zeigt sich, daß die hybriden Systeme basierend auf dem geteilten NN in einer verteilten Umgebung wesentlich flexibler als Gauß-Systeme sind. Bei Verwendung von Phonem-Modellen (3 HMM Zustände pro Phonem) kann das Netz je nach Anforderung ausgetauscht werden, während die HMM auf dem Server unverändert bleiben. Dazu sind zwei verschiedene NN erstellt worden, einmal mit 12 MFCC-Merkmalen, Energie und Zeitableitungen und einmal mit 9 RASTA-PLP Merkmalen, Energie und Ableitungen. Die RASTA-PLP [5] Analyse ist deutlich robuster gegenüber Hintergrundgeräusch und Kanalveränderungen. Insgesamt liefern Ganzwortmodelle für diese Aufgabe die besten Ergebnisse, sind allerdings schlecht auf größere Aufgaben übertragbar sind. Das hybride Ganzwortmodell mit einem RASTA-PLP Client übertrifft alle anderen Systeme. Phonemmodelle sind dem gegenüber universell einsetzbar, allerdings zeigt sich ein Qualitätsverlust. Ein Austausch des Clients durch Verwendung von RASTA-PLP Merkmalen anstelle von MFCC im hybriden Phonemsystem ergibt eine deutliche Verbesserung, ohne eine Veränderung am Server vornehmen zu müssen. Da bei einem Gauß-System die Modelle auf dem Server von den Merkmalen abhängen, ist hier ein solcher Austausch nicht möglich.

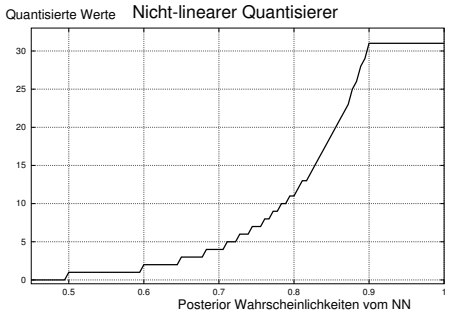


Abbildung 3: Skalarer Quantisierer für NN-Ausgänge

Zusammenfassung

Akustische Modelle in der verteilte Spracherkennung unterliegen teilweise anderen Anforderungen, wie sie an ein Standardsystem gestellt werden. Wir haben das Verhalten von etablierten Gauß-HMM unserer hybriden Architektur in einem Client/Server Szenarion gegenübergestellt und unter Benutzung der AURORA2-Datenbasis evaluiert. Bei Verwendung der hybriden Architektur ist ein Austausch des Clients möglich, ohne den Server zu verändern, womit die Schächen des Phonemmodells durch einen besser geeigneten Client aufgefangen werden können. Ebenson erlaubt die hybride Architektur auf dem Server die HMM-Topologie unabhängig vom Client zu verändern, sofern die Anzahl der übertragenen Wahrscheinlichkeiten festgehalten wird. Ein hybrides System unter Verwendung von RASTA-PLP Merkmalen hat das beste Erkennungsergebnis erzielt.

Literatur

[1] "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms," in *ETSI ES 201 108 v1.1.1 (2000-02)*, 2000.

[2] Herve Bourlard and Nelson Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, 1994.

[3] Jan Stadermann and Gerhard Rigoll, "Flexible Feature Extraction and HMM Design for a Hybrid Distributed Speech Recognition System in Noisy Environments," in *IEEE Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hongkong, China, Apr. 2003.

[4] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, 2000.

[5] Hynek Hermansky and Nelson Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

System	**A**	**B**	**C**
Ganzwort Gauß-HMM	13,3	15,6	18,0
Ganzwort NN/HMM (MFCC)	9,4	17,7	19,9
Ganzwort NN/HMM (RASTA)	10,6	13,5	13.1
Phonem Gauß-HMM	16,8	22,7	23,4
Phonem NN/HMM (MFCC)	14,3	25,9	27,4
Phonem NN/HMM (RASTA)	13,1	17,4	16.3

Tabelle 1: Wortfehlerraten (%) für verschiedene verteilte Systeme mit den AURORA2 Tests