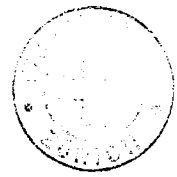
Diss. ETH No. 11662

Comparison of Classic and Hybrid HMM Approaches to Speech Recognition over Telephone Lines

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Doctor of Technical Sciences

presented by
HANS-PETER HUTTER
Dipl. El.-Ing. ETH
born 13 September 1960
citizen of Oberriet SG



accepted on the recommendation of
Prof. emer. Dr. W. Guggenbühl, examiner
Prof. Dr. A. Kündig, co-examiner

1996

Abstract

The subject of the present dissertation is the automatic speaker-independent recognition of isolated German digits spoken (by Swiss people) over the public switched telephone network. The approaches considered for this task are all based on hidden Markov models (HMMs). In addition to the classic HMM approaches, several connectionist ideas are investigated in order to improve the discrimination capability of the classic HMM systems.

In the first part, the theoretical foundations of HMMs in general and discrete density HMMs (DDHMMs) in particular are summarized. After that, several connectionist ideas, also termed hybrid HMM systems in the sequel, are theoretically discussed. In particular, a so-called *connectionist-SCHMM approach based on classic semi-continuous* HMMs (SCHMMs), which has been proposed by the author, is introduced.

The second part goes into the details of the design of the experimental system RECO, which is based on DDHMMs. It covers the speech data collection for the training and test data as well as the different experiments that have led to the high performance of the recognizer. With 98.6 % word recognition rate, the resultant DDHMM recognizer yields the same performance as the COST 232 reference recognizer based on continuous density HMMs on the same speaker-independent test set.

Different connectionist ideas of the first part are compared with each other and with classic HMM approaches in the last part of the thesis. The comparison shows, on the one hand, that the connectionist-SCHMM system proposed exhibits the best performance of all hybrid

HMM approaches followed by a learning vector quantization (LVQ3) approach. The connectionist-SCHMM recognizer uses a multilayer Perceptron (MLP) with three vectors input context as a posteriori probability estimator of phonetic element classes. The investigations confirm, on the other hand, that these connectionist ideas can really augment the discrimination capabilities of classic HMM systems, as they lead to significant improvements to the recognition rate of the classic HMM approaches for a single feature, i. e., the weighted LPC cepstrum.

Different concepts are investigated in order to incorporate multiple features into the connectionist-SCHMM system. The best concept has proven to be the use of separate MLPs for the weighted LPC cepstrum, the delta cepstrum, and the combination of log-energy and delta energy.

With this concepts, the connectionist-SCHMM recognizer reveals, as well as the LVQ3 approach, a 30 % reduced error rate compared to the classic HMM systems investigated on the two features weighted LPC cepstrum and delta cepstrum.

When log-energy and delta energy features are added, the performance of the baseline SCHMM approach is still outperformed by the LVQ3 and the connectionist-SCHMM approaches. For the DDHMM system, the modeling assumptions seem to be better met with all four features than with only two of them, resulting in a similar performance compared to the hybrid HMM systems. The connectionist-SCHMM and the DDHMM system, however, have only very few errors in common, so that a combination of the two recognizers entails an impressive error reduction of either system.

Kurzfassung

Das Thema der vorliegenden Dissertation ist die automatische sprecherunabhängige Erkennung von isolierten deutschen Ziffern gesprochen (von Schweizern) über das öffentliche Telephonnetz. Die untersuchten Ansätze für diese Erkennungsaufgabe basieren alle auf Hidden-Markov-Modellen (HMM). Neben den klassischen HMM-Ansätzen werden verschiedene konnektionistische Ideen untersucht mit dem Ziel, die Diskriminationsfähigkeit der klassischen HMM zu verbessern.

Im ersten Teil werden die theoretischen Grundlagen der HMM im allgemeinen und der diskreten HMM (DDHMM) im speziellen zusammengefasst. Danach werden verschiedene konnektionistische Ideen, im folgenden auch hybride HMM-Ansätze genannt, theoretisch diskutiert. Im besonderen wird der vom Autor vorgeschlagene, sogenannte konnektionistische SKHMM-Ansatz, der auf klassischen semikontinuierlichen HMM (SKHMM) basiert, näher vorgestellt.

Der zweite Teil erklärt im Detail den Aufbau des Experimentalsystems RECO, das auf DDHMM basiert. Dies umfasst sowohl die Aufnahme des Sprachmaterials für die Trainings- und Testkorpora als auch verschiedene Experimente, die zur hohen Erkennungssicherheit des Systems geführt haben. Mit 98.6 % liefert der DDHMM-Erkenner die gleich hohe Worterkennungsrate bezüglich des sprechenunabhängigen Testsets wie der COST 232-Referenzerkenner, der auf kontinuierlichen HMM basiert.

Verschiedene der im ersten Teil beschriebenen konnektionistischen Ideen werden im letzten Teil der Dissertation miteinander und mit den klassischen HMM-Ansätzen verglichen. Der Verleich zeigt auf der einen

Seite, dass das vorgeschlagene konnektionistische SKHMM-System die höchste Erkennungrate aller hybriden HMM-Ansätze aufweist gefolgt von einem Learning-Vector-Quantization-Ansatz (LVQ3). Der konnektionistische SKHMM-Ansatz verwendet dabei ein Mehrschichtperzeptron (MSP) mit einem Kontext von drei Vektoren am Eingang, um die A-posteriori-Wahrscheinlichkeiten der phonetischen Elementklassen zu schätzen. Die Untersuchungen bestätigen andererseits, dass die Diskriminationsfähigkeit von klassischen HMM-Systemen tatsächlich mit diesen konnektionistischen Ansätzen verbessert werden kann, indem diese zu einer signifikanten Verbesserung der Erkennungsrate der klassischen HMM-Ansätze bei Verwendung eines einzelnen Merkmals, des gewichteten LPC-Cepstrums, geführt haben.

Im Zusammenhang mit dem konnektionistischen SKHMM-System werden verschiedene Konzepte für den Einbezug mehrerer Merkmale untersucht. Als beste Variante stellt sich dabei die Verwendung von je einem separaten MSP für das gewichtete LPC-Cepstrum, das Delta-Cepstrum, sowie für die Kombination von Log-Energie und Delta-Energie heraus.

Mit diesem Konzept erreicht der konnektionistische Ansatz, gleich wie der LVQ3-Ansatz, eine um 30 % verminderte Wortfehlerrate verglichen mit den klassischen HMM-Systemen bei der Verwendung der zwei Merkmale gewichtetes LPC-Cepstrum und Delta-Cepstrum.

Beim zusätzlichen Einbezug der Merkmale Log-Energie und Delta-Energie schneiden die hybriden HMM-Ansätze gegenüber den klassischen SKHMM nach wie vor besser ab. Für das DDHMM-System scheinen die Modellannahmen bei der Verwendung von vier Merkmalen besser mit der Realität übereinzustimmen als mit zwei Merkmalen, was sich in einer praktisch gleich hohen Erkennungsrate wie die der hybriden HMM-Systeme äussert. Da der DDHMM-Erkenner und das konnektionistische SKHMM-System jedoch nur sehr wenige ihrer Erkennungsfehler gemeinsam haben, kann durch die Kombination der beiden Systeme die Fehlerrate jedes einzelnen nochmals markant reduziert werden.