A Two-Level Classification Scheme for CDHMM-Based Discrete-Utterance Recognition

Jan NOUZA
Department of Electrical Engineering, Technical University of Liberec
461 17 Liberec, Czech republic
tel. +42-48-254 41, e-mail: jan.nouza@vslib.cz

Abstract

In the paper a method or speeding up the response of a CDHMM based speech recognition system is introduced. The method, applicable for the recognition of discrete utterances, uses a two-level classification scheme. It consists in a fast match done with simplified models, followed by a final accurate match with a limited number of selected standard models. In this way the recognition time can be reduced by great deal without any significant loss of recognition accuracy. The method has been successfully applied in the design of real-time speech recognition systems operating with small and middle-size vocabularies.

Keywords:

speech recognition, hidden Markov model, continuous density HMM

1. Introduction

During the several last years, the hidden Markov model (HMM) technique has been playing the dominant role in the speech recognition area. The reason is that the statistical approach used in the HMM method seems to be more appropriate for speech analysis rather than some other identification techniques.

The HMM method offers two main advantages. First, it is capable of representing basic speech objects by simple probabilistic models that - if trained on a large multi-speaker material - may become speaker independent. Second, the same statistical approach can be applied both for the classification of words and subword units (like phonemes) as well as for modelling higher language structures (like phrases).

There are two basic versions of the HMM - a discrete HMM and a continuous density HMM (CDHMM) [1]. While the former operates with discrete symbols, which is made possible by vector quantisation of signal parameters, the latter uses continuous probability density functions to model the distribution of the parameters. It is evident that the latter approach enables finer modelling and at the same time it eliminates errors inherent in the vector quantisation. Thus in practice, CDHMM systems achieve higher recognition accuracy compared with the discrete HMM ones.

On the other side, one of the most serious drawbacks of the CDHMM technique is its very high computational complexity. This can cause troubles if we want to use the technique in a real-time speech recognition system. Methods of speeding up the recognition process are therefore intensively searched. One possible solution, applicable to a discrete-utterance recognition task, is described in this paper.

2. The discrete-utterance recognition using CDHMM

The discrete-utterance recognition (DUR) is a simplified subtask of the general speech recognition problem. Within the DUR, an utterance (either a single word or a word sequence) is spoken separately, i.e. it is preceded and followed by short pauses. Each vocabulary utterance has its model in the recognition system. The system classifies an unknown utterance by matching its parametric representation with all the models in order to find the most likely match.

Let us suppose that the utterance to be classified as one of N vocabulary items is represented by multidimensional vector $\mathbf{X} = (\mathbf{x}_1, \ldots \mathbf{x}_f, \ldots \mathbf{x}_F)$ consisting of F frame vectors $\mathbf{x}$, each being composed of P speech parameters $\mathbf{x} = (x_1, \ldots x_p, \ldots x_P)$. For modelling an utterance we employ the hidden Markov model form depicted in Fig.1. Such a model consists of S states between which transitions are allowed only from right to left with given probabilities a_{ij}. Transition probability a_{ij} is defined as follows:

$$a_{ij} = \text{Prob}\,(q_t = Q_j | q_{t-1} = Q_i) \qquad (1)$$

The output function b_s associated with state s is, in the case of the CDHMM, a continuous probability density function made up as a mixture of M normal distributions:

$$b_s(\mathbf{x}) = \sum_{m=1}^{M} \frac{w_{sm}}{\sqrt{(2\pi)^P \det \mathbf{C}_{sm}}} . \exp[-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_{sm})^T \mathbf{C}_{sm}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_{sm})] \quad (2)$$

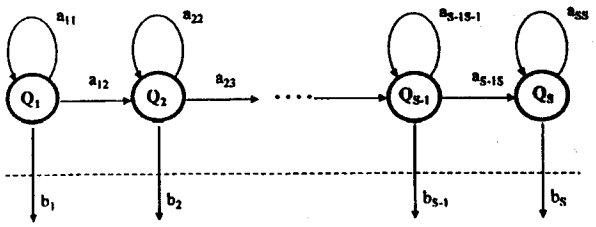

Fig.1 The HMM topology used for discrete utterances

In most practical DUR systems, models belonging to individual utterances have the same number of states. The utterance-specific characteristics are hidden in corresponding models in the values of transition probabilities a_{ij} and in the output function parameters $\bar{\mathbf{x}}_{sm}$ (the mean), $\mathbf{C}_{sm}$ (the covariance matrix) and w_{sm} (the mixture-weighting factor). All these parameters must be estimated during a system training procedure.

Within the classification process, a measure of similarity between the unknown utterance and each of the models is evaluated. The measure has a meaning of the log likelihood that the utterance is generated by a given Markov model. The likelihood of model Ψ is evaluated for the most probable state sequence Q^* using the Viterbi algorithm [1]:

$$\ln P(\mathbf{X}|\Psi) = \sum_{f=1}^{F} (\ln a_{q_{f-1}q_f} + \ln b_{q_f}), \quad (3)$$

where $q_f \in Q^*$

This is done repeatedly for each of the N models representing the complete vocabulary. The model n^* that matches best the utterance, i.e. that fulfilling equation

$$n^* = \arg \max_{n=1..N} \ln P(\mathbf{X}|\Psi_n) \quad (4)$$

determines the result of the classification.

In our previous study [2] we showed that the recognition rate of a CDHMM system is essentially a function of the P (the number of speech parameters), S (the number of states) and M (the number of mixtures). Also the recognition time depends on variables P, S and M, and further on the size of the vocabulary, i.e. on the number of models N. It was observed, however, that while the recognition time increases nearly proportionally with increasing values of these parameters, the recognition rate becomes saturated for certain values P_0, S_0 and M_0. These values, that can be found for the given system experimentally, represent the optimal settings both from the system accuracy and response time points of view. Any attempt to reduce the recognition time by decreasing the parameters below these optimal values will result in a worse performance of the recognition system.

3. A two-level classification approach

An alternative way to speed up the classification without loosing the accuracy exists. It consists in employing a two-level classification approach, which, particularly, in connection with the CDHMM technique may bring a considerable computation yield.

As mentioned above, the time T needed for the standard-scheme classification of one utterance depends on the system's parameters as follows:

$$T = f(N,P,S,M) = N.g(P,S,M) \quad (5)$$

where function g is a nonlinear function of P, S and M.

A two-level classification scheme assumes two sets of models: A set of standard models trained with parameters P_A, S_A and M_A, which are set equal to the optimal parameter values P_0, S_0 and M_0, and another set of simplified models trained with parameters $P_F < P_0$, $S_F < S_0$ and $M_F < M_0$. The classification consists in two steps. Within the first one, a *fast match* between the unknown utterance and the simplified models is performed and the models are ordered according to the achieved likelihood scores. In the second *accurate match*, only the first N_A models (the standard ones) are used. The time consumed by the two matches will be given by the following equation:

$$T_2 = f(N,P_F,S_F,M_F) + f(N_A,P_A,S_A,M_A) \quad (6)$$

The time reduction factor, defined as the ratio between time T_2 needed for the two-level scheme and time T_1 of the one-level scheme can be derived from eq. (5) and (6):

$$\frac{T_2}{T_1} = \frac{g(P_F,S_F,M_F)}{g(P_A,S_A,M_A)} + \frac{N_A}{N} \quad (7)$$

From equation (7) we may observe that the factor will depend on the fast match parameters P_F, S_F, M_F and on the number of models N_A selected for the accurate match. In order to find the minimum of eq. (7) for the given system, a set of experiments with various values of P_F, S_F, and M_F must be run. As we show further, the number N_A depends on these three parameters.

The parameter M_F can be set equal to 1 because in practice many CDHMM systems operate succesfully in a single-mixture mode. In this specific case, function g can be approximated in the following way:

$$g(P,S,1) = k_0 + k_P.P + k_S.S + k_{SP}.S.P \quad (8)$$

with constants k_0, k_S, k_P and k_{SP} being dependent on the given computing system. The formula (8) may be of practical use, namely, for the fast estimation of the reduction factor given by equation (7).

The parameters P_F and S_F have to be searched experimentally. Since in practical systems the S_0 value lies in range 6 - 14, the optimal value of the S_F should be searched in range 3 to $S_0/2$. The choice of the P_F is a more difficult task. Actually it is not only an issue of the number, but it is a question which and how many parameters should be chosen. This leads to the classic feature selection/extraction problem. A method for ordering speech parameters according to their importance with respect to the CDHMM technique was presented, for example, in [3]. We have applied another approach (described in [4]) that gives even better results. Anyway, having the speech parameters ordered, we can simply focus on experimenting with different values of P_F.

The choice of the last parameter N_A is constrained by the request that the recognition rate R_2 of the two-level system should not be worse than that of the original one-level system, R_1. In practice, a certain minor decrease of the accuracy (for example, by $\varepsilon < 0.2\%$) is still acceptable. Thus, for each pair P_F and S_F we search for such N_A so that $R_2 > R_1 - \varepsilon$.

The complete design of a two-level classification system may proceed as follows: First, for the given application and the given testing database we find the P_0, S_0 and M_0 values such that the recognition rate achieves the maximum. Than, on the same database, we run a set of experiments with varying values of P_F and S_F, for which we find the corresponding N_A and evaluate eq. (7). The minimum of eq.(7) determines the optimal choice of parameters for the two-level system. This is illustrated in Fig.2 which is based on experiments done with the CAD_CZ database. The above described procedure can be automated and included into the training process of the recognition system.

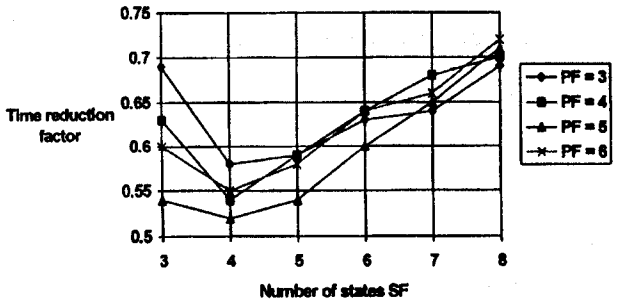

Fig.2 The time reduction factor as a function of the fast match parameters P_F and S_F (data from experiments done on CAD_CZ database)

4. Practical applications

The proposed method has been applied in the design of two practical speech recognition systems. In the first application, the vocabulary consists of 33 Czech words, in the second application the recognizer operates with 121 multi-word utterances (mostly Czech city names). Both the systems are speaker-independent and run in real time on a common personal computer (with processor 80486). Table 1 offers a comparison between the standard and the two-level classification scheme. We can see that the latter method reduces the classification time to 52% in the first case and to 29% in the second case. In both the cases the recognition accuracy remains almost unchanged. We may assume that for larger vocabularies the time reduction factor might be even more favourable.

Table 1 Comparison of standard and two-level classification schemes used in two practical speech recognition systems

Database name:	CAD_CZ	BUS
Vocabulary size:	33 words	121 utterances
Number of tested speakers	48	24
One-level classification scheme		
System parameters: $P_A/S_A/M_A$	18/8/1	18/14/1
Recognition rate [%]	98.36	96.91
Recognition time [ms]	320	1450
Two-level classification scheme		
System parameters: $P_F/S_F/M_F$	6/4/1	8/6/1
Recognition rate [%]	98.36	96.87
Recognition time [ms]	167	420

5. Conclusions

In the paper we have proposed a two-level classification scheme that is applicable for discrete-utterance recognition systems based on continuous density hidden Markov models. The scheme combines a fast match performed with simplified models and the accurate match limited to a small number of selected standard models. The construction of the fast match models is based on reducing numbers of model states and mixtures and, particularly, on selecting only the most important speech parameters. The above presented results of several practical tests demonstrate that the proposed method may lead to a considerable reduction of the classification time while the impact on the recognition accuracy is negligible.

6. References:

[1] HUANG, X.D., ARIKI,Y., JACK, M.A.: Hidden Markov Models for Speech Recog-nition. Edinburgh University Press, Edinburgh, 1990.

[2] NOUZA, J: Computation Speed Considerations in Speech Recognition Based on Conti-nuous HMMs. In Proc. of conference ELEKTRO'95, Žilina, February 1995, pp.57-60.

[3] BOCCHIERI, E.L., WILPON, J.G.: Disciminative Feature Selection for Speech Recognition. Computer Speech and Language, 1993, No.7, pp 229-246.

[4] NOUZA, J: On Speech Feature Selection Problem: Are Dynamic Parameters More Important than the Static Ones? In Proc. of EUROSPEECH'95 conference, Madrid, September 1995, (in print).

About author...

Jan NOUZA (38) is a senior lecturer at the department of electrical engineering of the Technical University of Liberec. He received his master's degree (Ing.) in 1981 and doctor's degree (CSc.) in 1986, both of them at the Czech Technical University in Prague. Dr. Nouza has been professionally interested in the speech processing domain for more than 15 years. His recent research is focused mainly on applications of hidden Markov modell techniques in speech recognition, on the development of speech dialogue systems and on speech aids for handicapped.

CONTENTS OF ELECTRICAL ENGINEERING JOURNAL

(Issued by the Faculty of Electrical Engineering and Informatics, Slovak Technical University, Bratislava).

(Only papers published in English are mentioned.)

No. 8 , 1995

PAPERS:

A Novel System for 3D Acoustic Object Recognition Based on the Modified Rapid Transform - *J. Turán, K. Althöfer*

KDV Solitons in a Nonlinear Transmission Line - *R. Kukuča, J. Oravec*

A Method of Pressure Determination in a Closed System - *J. Ivan, V. Dubravcová*

The Dataflow Computer Architecture with Direct Operands Matching - *M. Jelšina*

COMMUNICATIONS:

I-V Characteristics and their Temperature Dependence in $In_{0.53}Ga_{0.47}As$ and GaAs Collector Regions of Heavily Doped HBTs - *S.S. De et all*

The Analysis of Radar Signals Able to Decrease the Clutter Level of Surrounding Objects and Meteorological Particles - *-D. Juřík, A. Hrbáň*

Recognition Unit Connections for Determining the Direction of Incremental Sensor Shaft Rotation - *M. Milly*

No. 9 , 1995

PAPERS:

New Decompsition-Coordination Methods for Control of Complex Systems, Part 1: Theory - *D. Chmúrny, R. Chmúrny*

Electronic Structure of Radiation Defect Calculated with a Semiempirical Approach - *P. Ballo, P. Macko, L. Harmatha, D. Rajniak*

Image Compression Using Neural Networks - *M. Oravec, P. Podhradský*

COMMUNICATIONS:

Measurement of Amplitude Permeability on High Permeability Toroids - *P. Butvin, B. Butvinová, J. Novák*

Recognition of Binary Images by Moment Invariants - *M. Durný, I. Mokriš*

Resolver with a Linear Output Volt-Angle Characteristic - *L. Hruškovic*