3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A STUDY ON VITERBI BEST-FIRST SEARCH FOR ISOLATED WORD RECOGNITION USING DURATION-CONTROLLED HMM

Masaharu Katoh and Masaki Kohda

Faculty of Engineering, Yamagata University
Yonezawa-shi, Yamagata, 992 Japan

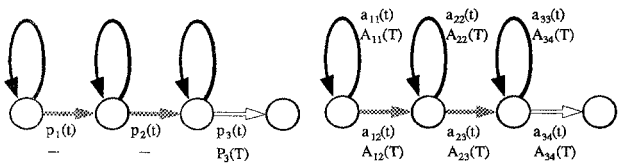
Abstract

In the conventional hidden Markov model (HMM), the probability of duration of a state decreases exponentially with time. It is not appropriate for representing the temporal structure of speech. To overcome this problem, the use of HMMs with duration models or time-dependent transition probabilities has been proposed [1][2]. These models accomplish the task with a large increase in the computation complexity. In this paper we present the Viterbi best-first searching algorithm using duration-controlled HMMs. To set a heuristic score appropriately, how the constraint is imposed on HMMs used in the backward Viterbi is investigated. The new searching algorithm is evaluated on isolated word recognition experiments. Experimental results show that the conventional Viterbi search with duration control takes 280~290 times of computation cost necessary for that without duration control, while the new searching algorithm takes only 10~15% of the computation cost keeping the same recognition rate.

1. Introduction

Hidden Markov models (HMMs) have been used successfully for speech recognition in the last several years. In the conventional HMM the probability of duration of a state decreases exponentially with time. This type of state duration probability does not represent the temporal structure of speech. To overcome this problem a number of alternatives have been proposed. One technique involves the use of HMMs with state duration models or time-dependent state transition probabilities [1][2]. These models accomplish the task with a large increase in the computation complexity. This is a major weakness in the use of duration-controlled HMMs for speech recognition.

In HMM-based speech recognition, an optimal path search by Viterbi algorithm is regarded as a problem of graph search. In this point of view the best-first searching algorithm can be applied to solve for the most likely path. Our previous paper [3] described the Viterbi best-first searching algorithm for isolated word recognition using the conventional HMM and showed that the searching algorithm can achieve a substantial reduction in computation complexity.



(a) Duration model (b) Time-dependent transition model

Fig. 1 Duration-controlled HMM.

In this paper, the above searching algorithm is extended to be applied to duration-controlled HMMs with the duration models of a state and a phoneme. A key point of the best-first searching algorithm is to estimate a heuristic score appropriately. The heuristic score which guarantees to find the most likely path while taking the less computation cost is desirable. To obtain the heuristic score the backward Viterbi algorithm is carried out. The HMM used in the backward process is produced beforehand on basis of the maximum output probability in the initial HMM.

Duration-controlled HMMs used in this paper are given in Section 2. The Viterbi best-first searching algorithm using the duration-controlled HMMs is proposed in Section 3. Isolated word recognition experiments for evaluating the proposed searching algorithm are mentioned in Section 4. Some experimental results are shown in Section 5.

2. Duration-controlled HMM

2.1 HMM with duration model [1]

A phoneme HMM with duration model is shown in Fig. 1(a). The transition to next state is determined stochastically by the probability $p_j(t)$ depending upon the duration in present state, t. When the duration in a phoneme is considered too, the transition to the last state of a phoneme model, i.e. to the first state of next

Table 1 Parametric distributions of duration model.

Geometric	$a_{jj}^{t-1}(1 - a_{jj})$	a_{jj} : transition probability
Uniform	$\frac{1}{t_1 - t_0 + 1}$	t_0 : minimum duration t_1 : maximum duration
Gaussian	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(t-\mu)^2}{2\sigma^2}}$	μ : average of duration σ^2 : variance of duration
Poisson	$\frac{\mu^t}{t!} e^{-\mu}$	μ : average of duration
Gamma	$\frac{1}{\Gamma(\alpha)\,\beta^\alpha} t^{\alpha-1} e^{-\frac{t}{\beta}}$	$\alpha = \frac{\mu^2}{\sigma^2}$, $\beta = \frac{\sigma^2}{\mu}$
Logarithmic Gaussian	$\frac{1}{\sqrt{2\pi\sigma'^2}\cdot t} e^{\frac{-(\log t - \mu')^2}{2\sigma'^2}}$	μ' : average of log-duration σ'^2 : variance of log-duration

phoneme model, is determined stochastically by the probability $P_J(T)$ depending upon the duration in present phoneme model, T.

As the duration model, six types of parametric distributions are investigated. They are given in Table 1.

2.2 HMM with time-dependent transition [2]

A phoneme HMM with time-dependent transition is shown in Fig. 1(b). The transition probability from state j to state k, given that the duration in the state j is t, is represented by $a_{jk}(t)$. When the duration in a phoneme is considered too, the transition probability from state j to state k, given that the duration in the phoneme model including state j is T, is represented by $A_{jk}(T)$. These transition probabilities satisfy the following equations for duration of any value.

$$a_{jj}(t)+a_{j\,j+1}(t) = 1$$
$$A_{11}(T) = A_{12}(T) = A_{22}(T) = A_{23}(T) = A_{33}(T)$$
$$A_{33}(T) + A_{34}(T) = 1$$

The actual transition is restricted so that the durations in a state and in a phoneme model will satisfy the following equations.

$$t_0 \le t \le t_1 \qquad T_0 \le T \le T_1$$

It should be noted that the time-dependent transition probabilities can be obtained from the distributions of duration model as shown below.

$$a_{j\,j+1}(t) = \frac{p_j(t)}{1 - \sum_{\tau=0}^{t-1} p_j(\tau)}, \qquad A_{34}(T) = \frac{P_3(T)}{1 - \sum_{\tau=0}^{T-1} P_3(\tau)}$$

A discrete HMM is considered. An output probability for the observation at input frame i given the transition from state j to state k is represented by $b_{jk}(O_i)$. A path score for the transition from state j to state k at input frame i, given the state duration t and the phoneme duration T, is obtained by the following equation.

$$s(t,T,i,j,k) = \log\{\, b_{jk}(O_i) \cdot a_{jk}(t) \cdot A_{jk}(T) \,\}$$

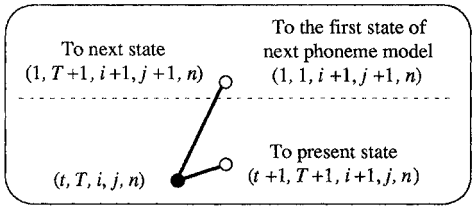


Fig. 2 Successor nodes expanded at node (t,T,i,j,n).

State duration	Phoneme duration					
	In present phoneme			To next phoneme		
	$T < T_0$	$T_0 \le T < T_1$	$T_1 = T$	$T < T_0$	$T_0 \le T < T_1$	$T_1 = T$
$t < t_0$						
$t_0 \le t < t_1$						
$t = t_1$						

Fig. 3 Successor nodes corresponding to the combination of state duration and phoneme duration.

3. Duration-controlled Viterbi best-first search

3.1 Outline of Viterbi best-first search

A situation that as the result of the transition at input frame i, stays in the j-th state of the n-th word HMM for duration t and in the phoneme model including its state for duration T is represented by node (t,T,i,j,n){ $t = 1 \sim t_1$, $T = 1 \sim T_1$, $i = 0 \sim I$, $j = 1 \sim J_n$, $n = 1 \sim N$ }. Successor nodes expanded at node (t,T,i,j,n) are shown in Fig. 2 : (t+1,T+1,i+1,j,n) for the transition to present state, (1,T+1,i+1,j+1,n) for the transition to next state, and (1,1,i+1,j+1,n) for the transition to the first state of next phoneme model. Considering the restriction on the actual transition, the successor nodes may be different according to the combination of the duration in a state and the duration in a phoneme model . They are shown in Fig. 3.

A path extension based upon the best-first search is made in accordance with the following procedure.

(a) Initialize open list **P** and closed list **Q**.

P = {(1,1,0,1,n) | n = 1 ~ N}

Q = NULL

P is the list for expandable nodes, and **Q** is the list for expanded nodes.

(b) Select the node having the maximum evaluation score from the list **P**.

(i) If the selected node is some terminal one $(1,1,I,J_n,n)$, the search process terminates with the recognition result that the input is word n.

(ii) Otherwise, the successor nodes expanded at the selected node are inserted into the list **P**, and the selected node is shifted to the list **Q**.

(c) Go to step (b).

The evaluation score at node (t,T,i,j,n) is given below

$$f(t,T,i,j,n) = g(t,T,i,j,n) + h(t,T,i,j,n)$$

where g(t,T,i,j,n) is the Viterbi score of the grown partial path from starting node to the node concerned, and h(t,T,i,j,n) is the heuristic score which represents the estimated score of the ungrown partial path from the node concerned to terminal node. As well known, if the heuristic score satisfies the admissible condition, the best-first search is equivalent to A* search.

3.2 HMMs for the heuristic score estimation

The heuristic score is precomputed by performing the Viterbi search directed from terminal node to starting node. Its accuracy depends on how the constraint is imposed on the HMM used in the backward Viterbi.

The paths extended by the backward Viterbi using three types of HMMs and by the forward Viterbi are illustrated in Fig. 4.

(1) Initial HMM

From training samples the HMM with 4 states and 3 loops is produced for each phoneme. In the backward Viterbi, the path score is defined by the output probability only . Transition probability is ignored. Thus the backward Viterbi score , i.e. the heuristic score, does not depend on duration parameters t and T. In other words, the backward Viterbi is performed without duration con-

trol.

The initial HMM is the same as the phoneme model used in the forward Viterbi best-first search. It is noted that in the best-first search, the path score is defined consistently by the output probability and the transition probability as described in Section 2.2.

(2) HMM based upon the maximum path score in a phoneme model

For each observation symbol, the maximum value of path scores, i.e. output probabilities, in the initial HMM is computed with the procedure shown in Fig. 5. Thus the 2-state/1-loop HMM having the above maximum value as path score can be produced for each phoneme.

(3) HMM based upon the maximum path score in a word model

From the word HMM obtained by concatenating the initial HMMs in accordance with the phonemic expression of each word, the maximum value of path scores in the word model is computed likewise. Thus the 2-state/1-loop word-HMM can be produced for each word. In this case the path score dose not depend on state number j, and depends on input frame i and word number n.

(4) HMM based upon the maximum path score through all the word models

From the word HMM obtained by concatenating the initial HMMs, the maximum value of path scores through all the word models is computed. Thus the 2-state/1-loop word-HMM having the same path score through all words can be produced. In this case the path score dose not depend on state number j and word number n, and depends on input frame i only.

4. Speaker-dependent word recognition experiments

ATR speech data including 5240 words and 216 phonetically balanced words uttered by one male speaker are used. The speech data is sampled at 12 kHz, pre-emphasized with a filter having a transfer function of $(1\text{-}z^{-1})$, and windowed using a 32 msec interval Hamming window every 8 msec. Then, LPC analysis is carried out, 16th-order LPC cepstral coefficients are obtained, and finally the VQ code sequence is generated. For VQ codebook generation, the speech data of 216 phonetically balanced words are used. The size of VQ codebook is 256.

The initial HMMs have been trained using a half of the speech data of 5240 words. In our experiments the HMM with time-dependent transition of Fig.1(b) is used as duration-controlled one. Its transition probability is obtained from duration model as described in Section 2.2. The parameters of distributions of Table 1 are obtained by performing the Viterbi algorithm with the same speech data as that for the HMM training, and dividing the speech data into segments corresponding to state and phoneme models.

To obtain the heuristic score, the following HMMs are used.

HMM-I: HMM based upon the maximum path score through all the word models.

HMM-II: HMM based upon the maximum path score in a word model.

HMM-III: HMM based upon the maximum path score in a phoneme model.

HMM-IV: Initial HMM.

The heuristic scores computed by performing the backward Viterbi with the above HMMs, together with the extremely over-estimated score of the value 0, are investigated. It is noted that these heuristic scores satisfy the admissible condition.

The computation complexity of the path extension is evaluated with the ratio of the number of nodes on the paths extended by the best-first search to that by the forward Viterbi. Likewise the computation complexity of the heuristic score estimation is evaluated with the ratio of the number of nodes on the paths extended by the

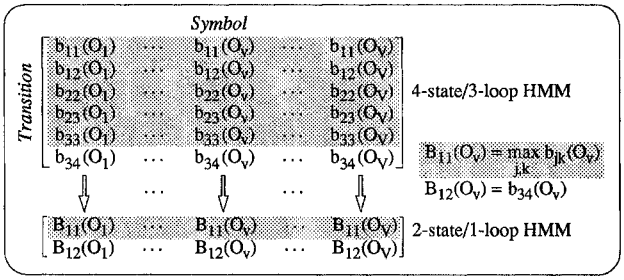
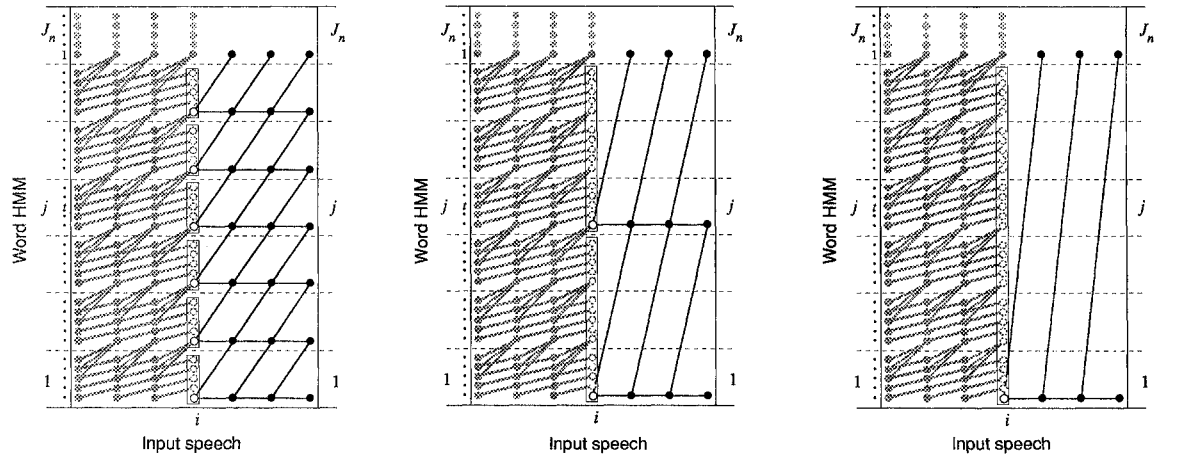


Fig. 5 Procedure of computing the maximum path score in a phoneme model.



(a) Initial HMM

(b) HMM based upon the maximum path score in a phoneme model

(c) HMM based upon the maximum path score in a word model

Fig. 4 The paths extended by the backward Viterbi using three types of HMMs and by the forward Viterbi.

backward Viterbi to that by the forward Viterbi. In these evaluations, the forward Viterbi means the conventional Viterbi search using the initial HMM without duration control.

5. Experimental results

The duration-controlled Viterbi best-first searching algorithm was evaluated on isolated word recognition experiments. Speech data of the first experiment was a small vocabulary of 216 phonetically balanced words. The experimental results of computation complexity are shown in Fig. 6(a). As can be seen from Fig. 6(a), the heuristic score using the HMM with the maximum path score in a word model has the smallest computation complexity.

Speech data of the second experiment was a large vocabulary of 2620 words which consist of a half of speech data of 5240 words and are different from that for the HMM training. In this experiment, the HMM with the maximum path score in a word model was used for the heuristic score estimation. The experimental results of recognition performance and computation complexity are shown in Fig. 6(b). The doted line in the right side of Fig.6(b) indicates the recognition rate by the conventional Viterbi search without duration control.

Fig. 7 shows examples of experimental results corresponding to the HMMs used to obtain the heuristic score. In Fig. 7, the sections from top to bottom show the number of expanded nodes, the number of word candidates, and the region of extended paths which belong to the trellis of the same category as input speech.

6. Conclusion

The Viterbi best-first searching algorithm was extended to be applied to duration-controlled HMMs. From the experimental results, the following were concluded.

(1) The HMM with time-dependent transition probability obtained from duration model was effective in attaining high recognition performance.

(2) In the Viterbi best-first search, the heuristic score using the HMM based upon the maximum path score was effective in reducing the computation cost significantly.

(3) The conventional Viterbi search with duration control took 280~290 times of computation cost necessary for that without duration control, while the duration-controlled Viterbi best-first searching algorithm took only 10-15% of the computation cost keeping the same recognition rate.

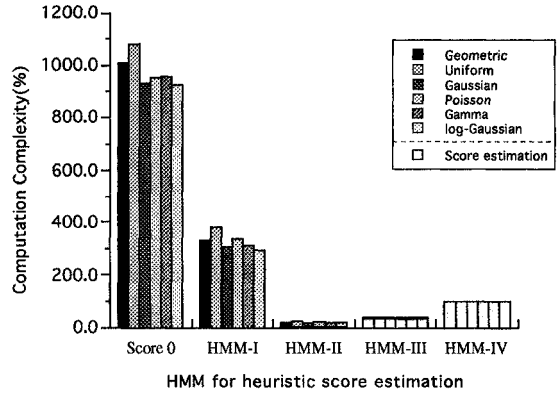
(4) With respect to duration model, the logarithmic Gaussian density was appropriate in a viewpoint of search efficiency.

References

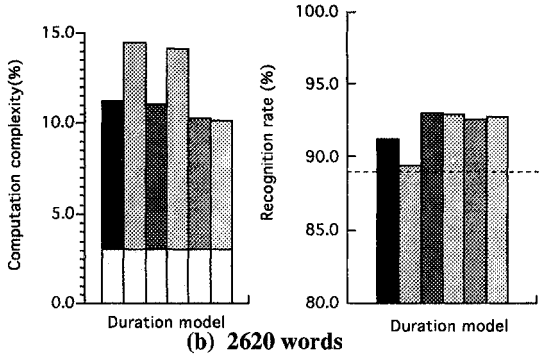
[1] J.D. Ferguson: "Variable duration models for speech", Proc. Symp. on Application of hidden Markov models to text and speech, pp.143-179 (Oct 1980).

[2] P. Ramesh, J.G. Wilpon: "Modeling state durations in hidden Markov models for automatic speech recognition", ICASSP92, Vol. 1, pp.381-384 (March 1992)

[3] M. Kohda, T. Kitamura: "A study on Viterbi best-first search for isolated word recognition based on discrete HMMs", Trans. of IEICE (D-II), Vol.J77-D-II, No.7 (July 1994).

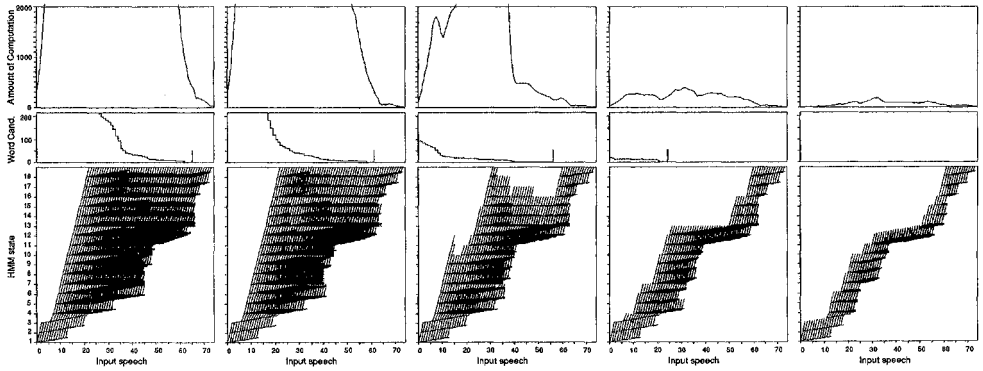


(a) 216 phonetically balanced words



(b) 2620 words

Fig. 6 Experimental results of isolated word recognition.



(a) Score 0 (b) HMM-I (c) HMM-II (d) HMM-III (e) HMM-IV

Fig. 7 Examples of experimental results corresponding to the HMMs used to obtain the heuristic score.