Proceedings of The Institute of Acoustics

IMPROVED DURATION MODELLING IN HIDDEN MARKOV MODELS USING SERIES-PARALLEL CONFIGURATIONS OF STATES

Anneliese E. Cook and Martin J. Russell.

Speech Research Unit, R.S.R.E. Malvern, Worcs.

INTRODUCTION

Hidden Markov models (HMMs) form the basis of many current speech recognition systems. Their advantage is that they provide a formal mathematical framework for modelling the variability in speech signals which is amenable to a set of computationally efficient and mathematically rigorous algorithms for automatic model parameter reestimation and pattern classification. However, several properties of HMMs are inappropriate in the context of speech pattern modelling; one such property is the underlying model of state duration in a HMM. This paper proposes one method of overcoming this problem by extending the HMM to give a more realistic durational structure to the model while still retaining its desirable mathematical properties. Results of isolated-word recognition experiments on digits and on a set of minimally distinct word pairs are presented.

HIDDEN MARKOV MODELS

A hidden Markov model comprises two related mechanisms; an underlying Markov chain, and an associated random function for each state in the chain. For the purposes of speech pattern modelling, the states of the underlying model should be thought of as the 'target sounds' which constitute a word; hence the state output functions model the statistical variation in the speech pattern, and the underlying Markov process models the temporal structure of the word. The underlying model is assumed to be first-order and finite, so that it is *memoryless* (i.e the state of the model at any time t is a function only of its state at the time instant $t-1$ and not of its past history).

An N-state hidden Markov model M may be completely specified by :-

1. An N-element initial state probability vector Π whose elements π_i are defined by
   π_i = P (state s_i at time 1)
2. An NxN transition probability matrix A whose elements a_{ij} are defined by
   a_{ij} = P (state s_j at time $t+1$ | state s_i at time t)
3. A set of N multivariate pdfs, $b_1,...b_i,...b_N$. Each b_i is a pdf defined on d-dimensional Euclidean space E^d such that for any vector v in E^d, $b_i(v)$ is the probability that v is generated by state s_i.

For the purposes of this paper, the output function b_i for each state in the chain is a multivariate Gaussian pdf parameterised by a mean vector m_i (representing the short-term spectrum) and diagonal covariance matrix v_i. The model must be entered through state 1 and left at state N.

Maximum likelihood classification using hidden Markov models.

Suppose that $O = (O_1,...O_t,...O_T)$ is a sequence of T vectors in E^d representing an unknown utterance. For the present paper, O_t should be thought of as a vector representing the short-term spectrum at time t during the utterance. The HMM, M, can only generate O via a state sequence $\sigma = \sigma(1)... \sigma(T)$ of length T (i.e. for each t, $\sigma(t)=s_i$ for some $i=1$ to N). The joint probability of O and σ conditioned on M is

$$P\ (\ O,\sigma\ |\ M\ ) = b_{\sigma(1)}(O_1)\prod_{t=2}^{T} a_{\sigma(t-1)\sigma(t)}b_{\sigma(t)}(O_t)$$

The probability P (O | M) of O given M is then

$$P (O \mid M) = \sum_{\sigma} P (O,\sigma \mid M)$$

where the summation is over all state sequences σ of length T.

Model Parameter Reestimation.

The first problem to be addressed when using HMMs for speech recognition is the creation of a model for each word in the vocabulary. The method used most frequently is to make an initial estimate of the model and improve it by an iterative procedure. Baum et al. [1], and Liporace [2] have shown how, given a HMM and a set of examples of the word, a new model can be derived such that the likelihood of the set of training words conditioned on the new model is greater than or equal to their likelihood conditioned on the original model. Repeated application of this process gives a hill–climbing algorithm which locally maximises P(O | M). The reestimation formulae used at each stage in the process are given below:–

$$\bar{a}_{ij} = \frac{\sum_{\sigma\epsilon S_{ij}} P(O,\sigma \mid M)}{\sum_{\sigma\epsilon S_i} P(O,\sigma \mid M)} \tag{1}$$

$$\bar{m}_i = \frac{\sum_t \sum_{\sigma\epsilon S_i(t)} P(O,\sigma \mid M)\, O_t}{\sum_{\sigma\epsilon S_i} P(O,\sigma \mid M)} \tag{2}$$

$$\bar{v}_i = \frac{\sum_t \sum_{\sigma\epsilon S_i(t)} P(O,\sigma \mid M)\, (O_t-\bar{m}_i)(O_t-\bar{m}_i)^*}{\sum_{\sigma\epsilon S_i} P(O,\sigma \mid M)} \tag{3}$$

where $S_{ij} = \{\sigma\colon \sigma(t)=s_i,\ \sigma(t+1)=s_j$ for some $t\}$, $S_i = \{\sigma\colon \sigma(t)=s_i$ for some $t\}$ and * denotes matrix transposition.

Computationally efficient algorithms exist for the evaluation of the reestimation formulae (1) to (3) [3]. In practice, extensions of these formulae which are able to accommodate several examples of a given word for parameter estimation are used.

DURATIONAL MODELLING IN MARKOV MODELS

The underlying model of state duration in a Markov chain is a geometric process. If the probability of a transition from a state to itself is a, then the probability $g(t)$ of remaining in that state for t time intervals is given by the *geometric* pdf

$$g(t) = a^{t-1}(1-a) \qquad t=1,2,\ldots \tag{4}$$

and the expected duration of the state is $(1-a)^{-1}$. This is an inappropriate model for the temporal variation in speech segments, as $g(t)$ decreases with increasing t; it is more likely that a segment of a word (as represented by a state in a HMM) will have a high probability of having some target duration, with lower probabilities assigned to longer and shorter durations [4].

One method of obtaining more realistic durational structures from Markov models is explicitly to associate a durational pdf with each state in the chain. The resulting model is known as a *Hidden Semi–Markov model* (HSMM). It has been shown that the standard HMM parameter reestimation and word classification techniques (as outlined above) can be extended to these HSMMs for certain classes of durational pdfs, including the Poisson pdf [5], [6], and the Γ–pdf

[7]. However, the method is computationally expensive.

The approach considered here is to identify each node in the underlying Markov chain with a network of states or *sub−HMM* with a single state output pdf. The motivation behind this approach is the fact that a rich family of pdfs can be obtained as the overall durational pdf of a series−parallel network of states in a Markov model. This method has previously been investigated by Mergel and Ney [8].

Theoretical Basis for the use of Sub−HMMs

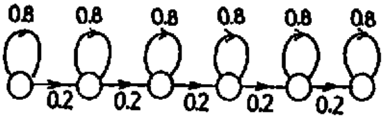
Cox [9] has shown that a durational pdf with a rational Laplace transform may be represented by a series−parallel network of exponential distributions whose lifetimes may be complex. The corresponding result for discrete time processes is that a duration pdf with rational *z−transform* may be represented as a series−parallel network of *geometric* processes. The proof of this result relies on a particular decomposition by partial fractions of the z−transform of the pdf; the parameters of the geometric pdfs in the equivalent network are derived from the roots of the polynomials which form the numerator and denominator of this z−transform. In general, these roots may be complex, and this gives rise to complex parameters in the corresponding geometric pdfs; the implications of this for speech signal modelling are not clear. So, in order to exploit Cox's results, further restrictions must be placed on the class of pdfs used to ensure that the parameters of the resulting geometric pdfs lie between 0 and 1; the general solution to this problem is unknown.

In this paper, three classes of series−parallel state configurations are considered :−

A. A series sub−HMM with tied self−transition probabilities.
B. A series sub−HMM with self−transitions and exit transitions.
C. A two−state recursive sub−HMM.

Type A. Figure 1 shows the topology of a type A sub−HMM; note that $a_{ii}=a$ (0.8 here) for all i.

Fig. 1.



The overall durational pdf n of such a network is the convolution of the (geometric) pdfs of the N individual states [10]. Although the pdfs thus obtained seem potentially useful, they are limited by the fact that the minimum duration of each state in the chain is 1, thus giving a minimum possible duration in the network of N. This can be overcome by introducing the concept of states with minimum duration 0.

The durational pdf of such a state is the *modified* geometric pdf, defined by:−

$$g_0(t) = (1-a)a^t \qquad t=0,1,2... \tag{5}$$

with expected value $\bar{g}_0 = a(1-a)^{-1}$
The z−transform of the modified geometric pdf is given by:−

$$\bar{g}_0(z) = \frac{1-a}{1-az} \tag{6}$$

Now consider an N−state type A sub−HMM, with minimum substate duration zero; the z−transform of the durational pdf of such a model is the product of the z−transforms of the modified geometric pdfs, i.e. :−

$$\bar{n}_0(z) = \frac{(1-a)^N}{(1-az)^N} \tag{7}$$

This is the z–transform of a *modified negative binomial* pdf. Hence, by the uniqueness property of the z–transform, the durational pdf for a series of states with minimum duration 0 is the modified negative binomial pdf n_0. The mean and variance of this distribution are *N* times those of the modified geometric, i.e.

$$\bar{n}_0 = \frac{Na}{(1-a)} \qquad \Delta n_0 = \frac{Na^2}{(1-a)^2}$$

Figures 2 and 3 illustrate the behaviour of the modified negative binomial distribution under certain constraints.

The modified negative binomial pdf has previously been used by Crystal and House [4], who investigated phoneme duration in American English. They found that such a pdf fitted the observed statistics very well. A further consideration is the fact that if the number of states *N* in the model is increased while holding the mean, $Na/(1-a)$, constant, the modified negative binomial pdf tends to a Poisson pdf (see fig. 2). This provides a link with the work described in [5] and [6] on HSMMs with Poisson state duration pdfs.

Fig. 2.
Modified negative binomial distribution with constant mean, 20, and 1,2,4,8,16,32,64 states, compared with Poisson distribution (solid line).

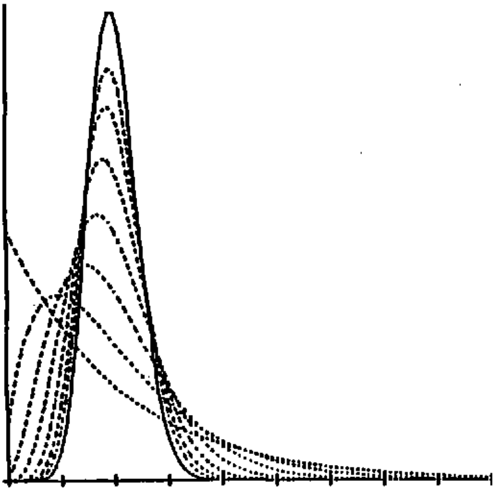
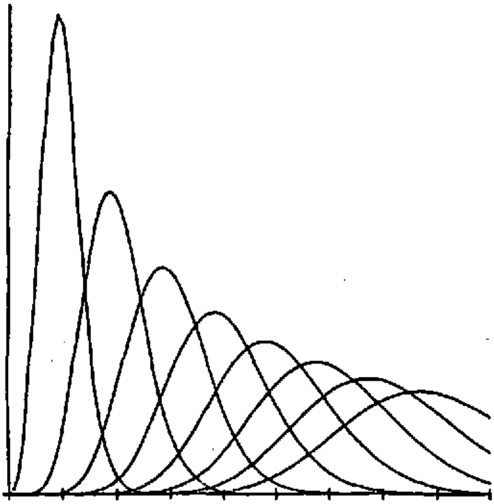


Fig. 3.
Modified negative binomial distribution with constant number of states, 32, & means 10,20,30,40,50,60,70 & 80.



States with minimum duration 0 are equivalent to the *null transitions* described in [11]. This property can easily be accommodated in the standard HMM parameter reestimation and pattern recognition algorithms.

<u>Type B.</u> Sub–HMMs of type B are standard HMMs in that the minimum duration of each substate is 1 and substate duration is modelled by a geometric pdf. The problem of increasing

minimum duration as the number of states in the sub−HMM increases is solved by allowing transitions from each substate to the final substate. In this model, the substate transition probabilities are not tied together during reestimation. Intuitively, in an N−state type B model the N exit transitions characterise the probabilities of durations 1 to N, and the self−transitions smooth the resulting pdf and extend it to infinity. Fig. 4 shows the topology and a typical durational pdf for a sub−HMM of type B.

Type C. In this class of sub−HMM each substate has minimum duration 1, and substate transition probabilities are reestimated independently. The potentially useful durational pdfs which arise with this type of model are a consequence of the recursive nature of the underlying sub−HMM topology. Figure 5 shows the topology and a typical durational pdf of a type C sub−HMM.

Fig. 4. Model B.

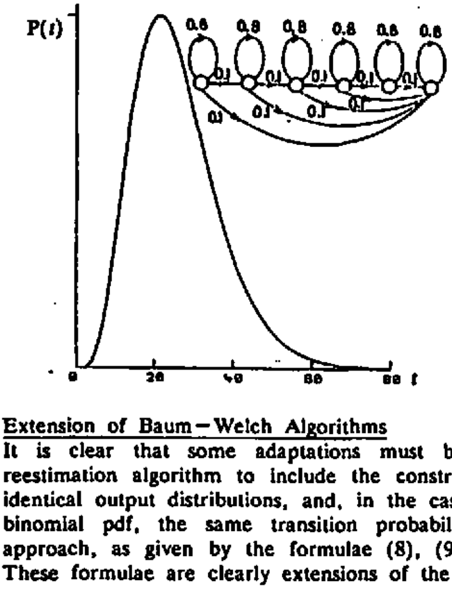
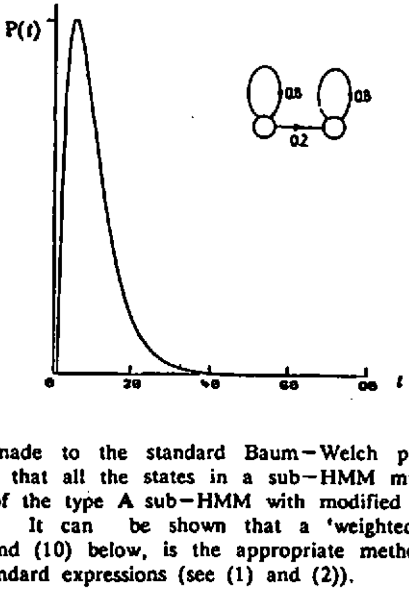


Fig. 5. Model C.



Extension of Baum−Welch Algorithms

It is clear that some adaptations must be made to the standard Baum−Welch parameter reestimation algorithm to include the constraint that all the states in a sub−HMM must have identical output distributions, and, in the case of the type A sub−HMM with modified negative binomial pdf, the same transition probability. It can be shown that a 'weighted mean' approach, as given by the formulae (8), (9) and (10) below, is the appropriate method here. These formulae are clearly extensions of the standard expressions (see (1) and (2)).

Suppose that each state s_i (henceforth referred to as a *macrostate*) in the N−state HMM $\mathbf{M}$ is expanded into an s−state sub−HMM $\mathbf{M}_i$ ($i=1,\ldots N$). This results in an expanded model $\mathbf{M}^*$ with Ns states. Let I_i denote the index of the first state in $\mathbf{M}^*$ which belongs to the sub−HMM $\mathbf{M}_i$, so that $I_i=(i-1)s+1$ ($i=1,\ldots N$).

In this case the modified reestimation formulae for the mean $\bar{m}_i$ and covariance matrix $\bar{v}_i$ of each state k in the i^{th} sub−HMM are given by :−

$$\bar{m}_k = \frac{\sum_t \sum_{j=I_i}^{I_{i+1}-1} \sum_{\sigma\in S_j(t)} P(O,\sigma \mid M)\, O_t}{\sum_{j=I_i}^{I_{i+1}-1} \sum_{\sigma\in S_j} P(O,\sigma \mid M)} \tag{8}$$

$$\bar{v}_k = \frac{\sum_{t} \sum_{j=I_i}^{I_{i+1}-1} \sum_{\sigma \in S_j(t)} P\,(O,\sigma \mid M)\,(O_t - \bar{m}_k)(O_t - \bar{m}_k)^*}{\sum_{j=I_i}^{I_{i+1}-1} \sum_{\sigma \in S_j} P\,(O,\sigma \mid M)} \qquad (9)$$

and, in the case of sub–HMMs of class A, the reestimation formula for the self–transition probability for each state k in the i^{th} sub–HMM is given by:–

$$\bar{a}_{kk} = \frac{\sum_{j=I_i}^{I_{i+1}-1} \sum_{\sigma \in S_{jj}} P\,(O,\sigma \mid M)}{\sum_{j=I_i}^{I_{i+1}-1} \sum_{\sigma \in S_j} P\,(O,\sigma \mid M)} \qquad (10)$$

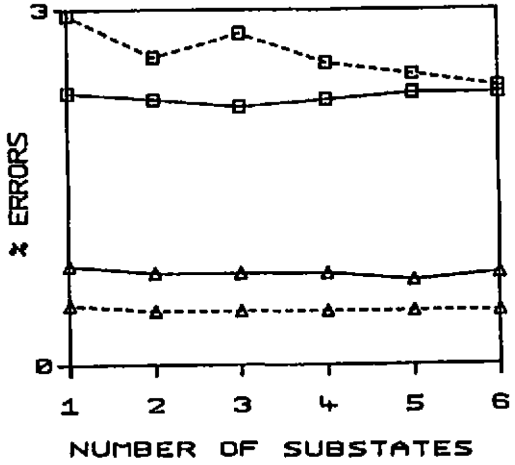
EXPERIMENTS ON SPEECH

Isolated Digit Recognition Experiments

The existence of a database of isolated digits (40 examples of each, spoken by each of 40 speakers [12]), added to the fact that digits would be widely used in potential applications of automatic speech recognition, was a prime motivating factor in the choice of vocabulary for these experiments. Results from a large–scale set of experiments on HMMs [13] using this database aided in the choice of parameters for the experiments; 10 examples of each word were used for training, and the models had 5– and 10– macrostates. Twenty of the speakers were selected; the three types of sub–HMM detailed earlier were used, and the number of states in the sub–HMMs was varied from 1 to 6.

Results.

Fig. 6. Percentage error rate vs. number of substates for digit recognition experiments.
—— = type A sub–HMM; □ = 5–macrostate model;
– – – – = type B sub–HMM; Δ = 10–macrostate model.



The results are summarised in figure 6. It may be seen that durational modelling does give a slight improvement in recognition accuracy for digits; this is most pronounced for type B sub–HMMs, where increasing the number of states gives a decrease in error rate from 3% for standard HMMs to 2.4% for 6–state sub–HMMs. Type C sub–HMMs give an error rate almost identical to 2–substate type B models, implying that the recursive property of the model has no particular advantage. Increasing the number of states in the sub–HMM did not have as significant an effect on recognition performance as increasing the number of macrostates in the model; this implies that the acoustic differences between digits are more significant than the durational differences.

Experiments on Minimally Distinct Word Pairs

The word pairs given below all have the property that their durational characteristics provide an important cue for discrimination [14]. For example, the voicing of the final consonant in 'league' extends the vowel duration over that in 'leek'; the initial fricative in 'seen' is longer than that in 'teen', although they are spectrally very similar.

chip/ship; cloze/close; five/fife; hard/heart; heard/hurt; league/leek; rider/writer; robe/rope; seen/teen; wand/want.

These 11 word pairs were used to test the performance of the modelling strategy; sub–HMM types A and B were used, with an 8–macrostate underlying model. Ten examples of each word were used for training and ten for recognition.

Results.

The results of these experiments are summarised in table 1; it is clear that durational modelling can improve recognition performance in this kind of task.

Table 1. Recognition errors in minimal pairs.

	Type A HMM No. of substates			Type B HMM No. of substates	
	1	4	8	1	4
Total Errors (%)	15.9	16.8	16.4	15.5	5.0

Sub–HMM type A gives no advantage over conventional HMMs; this is an unexpected result, in view of the relationship between the modified negative binomial and the Poisson pdfs, since experiments on the same data have shown that HSMMs with Poisson durational pdfs can achieve a 20% reduction in error rate [15].

Model B shows a significant decrease in error rate, similar to the improvement obtained using HSMMs with discrete duration pdfs [15]. This result is due to the ability of the type B topology to specify a tightly–limited durational pdf and hence to give more significance to durational differences.

CONCLUSIONS

The results presented here show that the performance of HMMs in speech recognition can be significantly improved by using appropriate state duration models. The relative simplicity and computational cheapness of the durational modelling technique presented here make it an attractive alternative to the use of semi–Markov models. The most effective type of sub–HMM appears to be the type B topology with self–transitions and exit transitions.

REFERENCES

[1] L.E. Baum, T. Petrie, G. Soules & N. Weiss, 'A maximisation technique occurring in the statistical analysis of probabilistic functions', Ann. Math. Stat., Vol. 41, no.1, 164-171, (1970).

[2] L. A. Liporace, 'Maximum likelihood estimation for multivariate observations of Markov sources', IEEE Trans. Inf. Theory, Vol. IT-28, no.5, 729-734, (1982).

[3] S. E. Levinson, L. R. Rabiner & M. M. Sondhi, 'An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition', Bell System Technical Journal, Vol. 62, no.4, 1035-1073, (1983).

[4] T. H. Crystal & A. S. House, 'Characterization and modeling of speech-segment durations', Proc. ICASSP 86, 2791-2794.

[5] M. J. Russell & R. K. Moore, 'Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition', Proc. ICASSP-85, Vol. 1, 5-8.

[6] H. Bourlard & C. J. Wellekens, 'Connected speech recognition by phonemic semi-Markov chains for state occupancy modelling', EUSIPCO-86.

[7] S. E. Levinson, 'Continuously variable duration hidden Markov models for speech analysis', Proc. ICASSP 86.

[8] D. Mergel & H. Ney, 'Phonetically guided clustering for isolated word recognition', Proc. ICASSP 85, Vol. 2, 854-857.

[9] D. R. Cox, 'A use of complex probabilities in the theory of stochastic processes', Proc. Cambridge Philosophical Soc., Vol. 51, Part 2, 313-319 (1955).

[10] K. S. Trivedi, 'Probability & statistics with reliability, queuing and computer science applications', Prentice-Hall, Ch. 2 (1982).

[11] F. Jelinek, R. L. Mercer & L. R. Bahl, 'Continuous speech recognition : statistical methods', Handbook of statistics (eds. P. R. Krishnaiah & L. N. Kanal), Vol. 2, 549-573 (North-Holland, 1982).

[12] D. C. Smith, M. J. Russell & M. J. Tomlinson, 'Rank ordering of subjects involved in the evaluation of automatic speech recognisers', RSRE memorandum no. 3926.

[13] M. J. Russell & A. E. Cook, 'Experiments in isolated digit recognition using hidden Markov models', this conference.

[14] M. J. Russell, R. K. Moore & M. J. Tomlinson, 'Some techniques for incorporating local timescale variability information into a dynamic time-warping algorithm for automatic speech recognition', Proc. ICASSP-83, Vol.3, 1037-1041.

[15] M. J. Russell, 'Experiments in isolated word recognition using hidden semi-Markov models', RSRE memorandum in preparation.