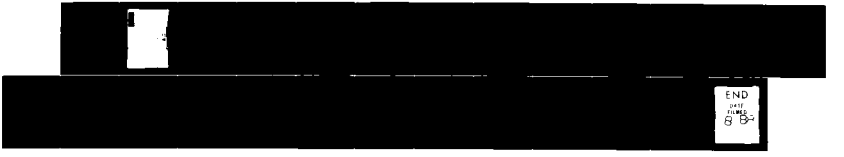
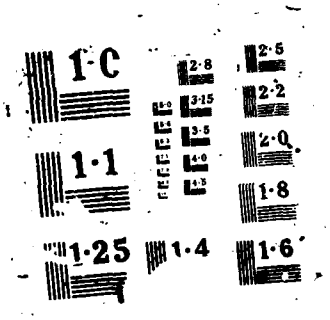
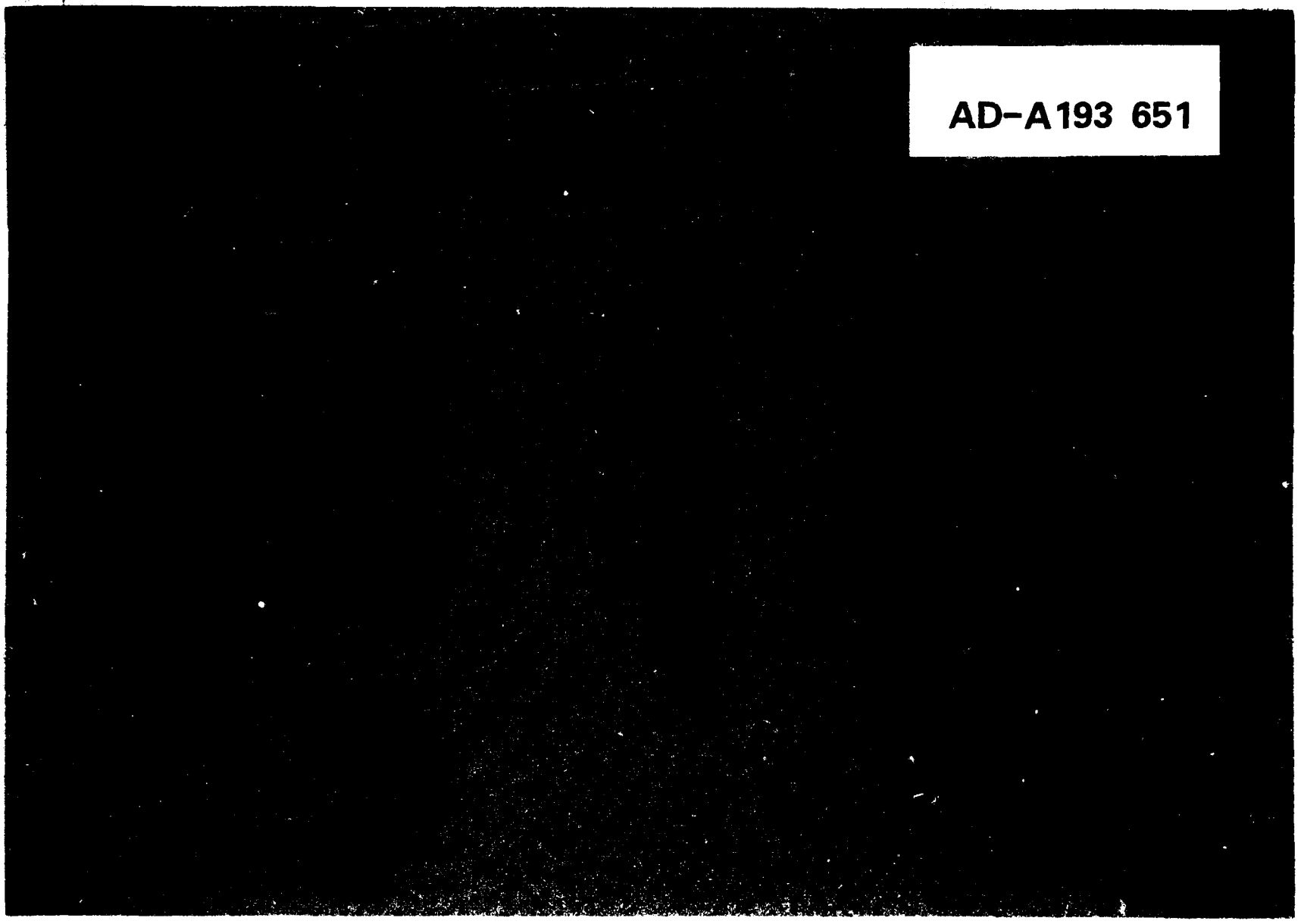
AD-A193 651 EXPERIMENTAL EVALUATION OF ALGORITHMS FOR CONNECTED SPEECH RECOGNITION US..(U) ROYAL SIGNALS AND RADAR ESTABLISHMENT MALVERN (ENGLAND) A E COOK NOV 87 1/1

UNCLASSIFIED RSRE-MEMO-4099 DRIC-BR-104991 F/G 12/3 NL



END
DATE FILMED
8 88


1·C
1·1
1·25
2·8
3·15
3·5
4·0
4·5
1·4
2·5
2·2
2·0
1·8
1·6


AD-A193 651

Royal Signals and Radar Establishment

Memorandum 4099

Experimental Evaluation of Algorithms for Connected Speech Recognition Using Hidden Markov Models.

Anneliese E. Cook

December 1, 1987

Abstract

Current Automatic Speech Recognition devices attempt to solve the connected word recognition problem by assuming that an unknown phrase is the output of a sequence of *statistical word-models. Typically, these models are constructed using examples* of words spoken in isolation; however, the acoustic patterns corresponding to words as they occur in fluent speech are quite different from those representing the same words spoken in isolation, and so the use in speech recognisers of models based on isolated utterances severely limits the performance of such devices. A method of extracting training utterances from fluent speech and constructing Hidden Markov Models (HMMs) from these templates, known as *Embedded Training*, is investigated here, in conjunction with a two-level algorithm for connected word recognition. The effects on recognition performance of various HMM training procedures are discussed, and experimental results are presented.

Contents

List of Figures

List of Tables

1 Introduction

Over the last few years, Hidden Markov Models (HMMs) have been shown to provide a powerful tool for automatic speech recognition [1]. Recent work at R.S.R.E. has concentrated on isolated word recognition using whole-word models, establishing a baseline performance criterion for standard parameter re-estimation and word recognition algorithms [2]; the present memorandum extends this investigation to connected word recognition.

Several dynamic programming algorithms for connected speech recognition are documented [3], [4], [5], [6], and one or two which make use of more heuristic approaches have been used with some success [7]. In HMM terms, all of these methods attempt to solve the connected word recognition problem by finding the sequence of HMM word-models which is most likely, in some sense, to have generated the unknown utterance. There are two closely related approaches to the solution of this problem. The first is a one-pass dynamic programming algorithm, as in [3], where the optimal sequence of HMMs and the state sequence which is most probable given this sequence of HMMs are computed simultaneously. The second approach involves two passes, one to find the HMMs which best represent each part of the unknown signal and one to find the optimal concatenation of the HMMs. The principal advantage of the two-level approach is that it can support a wide variety of methods for comparing HMMs with parts of the unknown signal; for example, the full likelihood calulations of the Baum-Welch recognition algorithm (which is necessary for some of the algorithms recently investigated at R.S.R.E. [13] for improving the modelling of duration in HMMs) require that the two- level recognition scheme be employed.

In the investigation of the connected word recognition algorithm, it was found that conventional, isolated-utterance training procedures were inadequate for dealing with fluent speech; they take no account of the coarticulatory effects between neighbouring words in rapid speech. Therefore, a method of using training utterances extracted from running speech, known as *embedded training*, and based on that described in [10], was developed to try and overcome this problem.

The connected word recognition algorithm was tested, using isolated word training and several variants of the embedded training strategy, on part of the NATO RSG10 speech database [9]. Triples and isolated examples of the English digits, spoken by male and female native and non-native English speakers, were used.

2 The Two-Level Connected Word Recognition Algorithm

In the two-level algorithm, the matching process is broken down into two main parts: the word-level decision, where the best explanation of every section of the unknown phrase is found *in terms of the HMM most likely to have generated it*, and the phrase-level decision, where the most probable partition of the phrase into these sections, and hence the most likely sequence of HMMs, is found.

2.1 Word-Level Matching

Let

$$\mathrm{O} = O_1, O_2, \ldots, O_t \ldots, O_T$$

represent an unknown phrase, of length T, so that O_t is a vector describing the acoustic properties of the signal at time t.

Let

$$\mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_n, \ldots, \mathrm{M}_N$$

denote HMMs representing the words in an N-word vocabulary.

Now define the *partial pattern*

$$\mathrm{O}_{l,m} = O_l, O_{l+1}, \ldots, O_m, \quad 1 \leq l \leq m \leq T$$

The first stage of the recognition process involves finding, for each l, m, the HMM $\mathrm{M}'_{l,m}$ which best explains the partial pattern $\mathrm{O}_{l,m}$. For example, if the classification criterion is maximum likelihood, then define

$$\mathrm{P}_{l,m} = \max_n(\mathrm{P}\,(\,\mathrm{O}_{l,m} \mid \mathrm{M}_n\,)) \quad \forall\, l, m \geq l$$

and hence

$$\mathrm{M}'_{l,m} = \arg(\max_n(\mathrm{P}\,(\,\mathrm{O}_{l,m} \mid \mathrm{M}_n\,))$$

2.2 Phrase-Level Matching

At this level, the partial explanations of the unknown signal computed in section 2.1 are used to find the sequence of HMMs which best explains the whole unknown signal. This is done by applying a dynamic programming (DP) algorithm to the partial explanations found at the previous level. The method involves the recursive application of dynamic programming to two matrices, denoted here by P and S. For each $l, m, m \geq l$, $P(l, m)$ is the probability of $\mathrm{O}_{1,m}$ given the best sequence of HMMs such that the final HMM in the sequence generates the sub- sequence $\mathrm{O}_{l,m}$. $S(l, m)$ is the time at which the penultimate HMM in the sequence is entered.

Initialise P and S by setting

$$P(1, m) = \mathrm{P}_{l,m}$$

and

$$S(1, m) = 0$$

The matrices are filled using the recursive DP algorithm outlined overleaf, where l and m denote the entry and exit times of the final HMM in the current sequence, and k denotes the entry time for the previous HMM.

For $l = 2, \ldots, T$

For $m = l, \ldots, T$

$$S(l,m) = \arg(\max_k(P(k,l-1)))$$

$$P(l,m) = \max_k(P(k,l-1).\mathrm{P}_{l,m})$$

Next m

Next l

In order to find the sequence of words, a backwards trace through S, the "previous model" matrix must be performed. First, the final model is identified by finding the entry point l of the HMM ending at time T which maximises the probability $\mathrm{P}_{l,T}$ at the final time instant; from this, the entry point and identity of the previous model are found from S and M′, and so the traceback continues until all the word boundaries have been found.

Note that in this scheme the number of words in the unknown utterance is not prespecified, but emerges when the most probable sequence of HMMs has been found.

Formally, let S_n represent the time at which the n^{th} HMM encountered during traceback is entered and E_n the instant when it is left. Then:-

$$\mathrm{S}_1 = \arg(\max_l(P(l,T)))$$

$\mathrm{E}_1 = T$

For $n = 2, n+1$, repeat until $\mathrm{S}_n = 1$:

$\mathrm{E}_n = \mathrm{S}_{n-1} - 1$

$\mathrm{S}_n = S(\mathrm{S}_{n-1}, \mathrm{E}_{n-1})$

Note that the optimal sequence of HMMs is decoded in reverse order; so if N HMMs were found to give the best explanation of the utterance then W_p, the identity of the p^{th} word in the utterance, corresponds to the HMM

$$\mathrm{M}'_{\mathrm{S}_{N-p+1},\mathrm{E}_{N-p+1}}$$

3 The Embedded Training Algorithm

Connected-word recognisers which use isolated word training procedures can work very well if the phrases to be recognised are spoken moderately slowly or if little coarticulation occurs between adjacent words. If the speech is faster, and a higher degree of coarticulation takes place, the models derived from isolated words will no longer provide an appropriate representation of the speech patterns. This places a severe limitation on the usefulness of such devices.

To overcome this problem, a method of extracting templates from running speech and creating HMMs based on these templates was devised, based on that reported in [10]. The method is outlined in the following sections.

3.1 Initial Model Estimation

First, a set of HMMs, one for each word in the vocabulary, is created from isolated utterances, to act as initial models. In the experiments described here, 10 isolated examples of each digit had all leading and trailing silences removed and were uniformly divided into 8 segments. Initital estimates of the mean and variance of the state acoustic vector and its expected duration were calculated from this segmentation. These initial model parameters were then re-estimated using the forward-backward (Baum-Welch) algorithm to give an optimal set of models from which to start the embedded training process.

3.2 Template Extraction–outer re-estimation loop

The models are aligned against known connected utterances using the Viterbi algorithm, to find the word boundaries. In the experiments described here, this alignment took the simplest form possible; the models for the words known to comprise the phrase were concatenated into one larger model which was matched against the string, and the connected phrase was segmented into its separate words according to the optimal state sequence through the models and the utterance.

When all the training phrases have been segmented, the optimal state sequences given by the Viterbi alignment are used to form new estimates of the mean, variance and expected duration of each acoustic state in all the models.

3.3 Model Optimisation–inner re-estimation loop

The full Baum-Welch re-estimation algorithm is used to re-estimate the model parameters in such a way as to find a local maximum of the likelihood of each HMM word-model over several iterations, using as training data the appropriate word-patterns extracted from the Viterbi alignment.

3.4 Optimisation Criteria

This alignment–re-estimation process (sections 3.2 and 3.3) is repeated until some optimisation criterion is reached. In the experiments described here, the criterion required that the mean log likelihood of the ten models should converge to a local maximum over successive alignment–re-estimation cycles, i.e. that from one outer loop to the next the mean log likelihood should increase until either a steady value was obtained or some maximum number of outer loop cycles had been completed.

4 Experiments

All experiments used the same data; English digits taken from the NATO RSG10 database [9]. Data from 17 people, male, female, native and non-native speakers of English, were used. For each speaker, the initial model building used 10 isolated examples of each digit; for the embedded training phase a list of 50 digit triples in which each digit occurred 15

times in a range of contexts, was used. The connected word recognition experiments were carried out on a total of 50 sets of 50 digit triples, up to four lists from each speaker.

The speech was processed using a 19-channel vocoder filterbank analyser with a 20ms frame rate [11].

The underlying structure of the HMMs was an 8-state model, with multivariate Gaussian states and diagonal covariance matrices. A form of amplitude normalisation was carried out on each frame of data; the mean over all channels was subtracted from each channel value and this value kept as an additional 20th channel. Two HMM topologies were used; the full upper-triangular model of which an example is shown in figure 1, with entry and exit permitted from any state, and the Bakis (tridiagonal) model shown in figure 2, which only allowed entry at state 1 and exit from state 8.

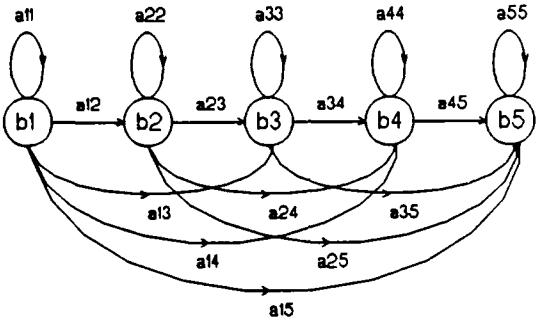


Figure 1: HMM with full upper-triangular transition probability matrix

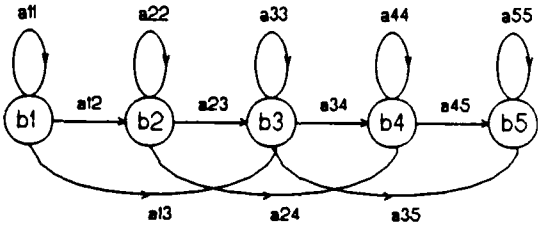


Figure 2: HMM with Bakis model transition probability matrix

4.1 Isolated Word Training

The models were created from isolated utterances. Ten examples of each of the ten digits were first uniformly segmented into 8 states, and the mean, variance and expected duration found for each state. The Forward-Backward algorithm was then applied to the models

to optimise the parameters. Two sets of experiments were carried out, one using the full upper-triangular transition probability matrix and one using the Bakis model.

4.2 Embedded Training

The models created in the experiments just described were used as the initial models for the experiments with embedded training; the training was performed using fifty digit triples, as described in section 4. Again, both model structures were used in the experiments.

4.2.1 Embedded training without Baum-Welch re-estimation

This experiment looked at the effect of omitting the "Inner Loop", as described in section 3.3, comparing the relative usefulness of the "Outer loop", template extraction part of the procedure and the model optimisation carried out in the "Inner Loop".

The models were formed by repeatedly aligning the composite model with the reference phrase and using the alignment to re-estimate the model parameters, again using a mean log likelihood criterion of convergence. It was noted during these experiments that, as one would expect, while the likelihood of individual models could increase or decrease at each step, the average log likelihood increased towards a maximum, and in almost every case the individual model likelihoods reached stable values.

4.2.2 Embedded training with durational constraints

The state duration model implicit in HMMs is a very poor model of the durational variability in speech sounds [12]; it favours short state durations, hence increasing the likelihood of insertion errors in connected speech recognition. Several methods of improving the state duration model have been proposed [13], [14]; however, the increase in computational cost of what is already a computationally intensive process motivated a search for a simpler form of duration modelling.

The method adopted used the expected word duration, rather than attempting to model state duration. At the final stage of the training procedure, the mean length and variance (assuming a Normal distribution) of each word in the vocabulary were calulated and stored as parameters of the model. At the word-level matching stage of the recognition process, the probability of the partial pattern conditional on the model was multiplied by a weighting factor (the probability of the model generating a pattern of that length). This method has previously been shown to be effective [15], and the increase in computational cost incurred is negligible.

5 Results

It should be noted that, as the data used in these experiments included a high proportion of speech from non-native English speakers, the results obtained will not be truly comparable with those quoted by other experimenters. For a valid comparison, table 1 includes in some of the results from the NATO RSG10 experiments [8].

PHRASE ERRORS OBTAINED ON ENGLISH DIGIT TRIPLES					
SYSTEM	NIPPON	VERBEX [1]	MOZART [1]	RSRE	HUMAN
% ERRORS	18.9	0.42	4.2	27	1.25

[1] *These recognisers used embedded training and therefore a reduced test-set.*

Table 1: Results of experiments carried out by RSG10

The results of all the experiments described in section 4 are shown in full in the Appendix; a brief summary is presented here.

ALGORITHM	% WORD ERRORS				% PHRASE ERRORS
	INSERTIONS	DELETIONS	SUBSTITUTIONS	TOTAL	
Isolated training, Full L-R	0.43	1.11	2.36	3.89	10.72
Isolated training, Bakis model	0.40	1.08	2.27	3.75	10.12
Embedded training, Bakis model	0.38	0.22	0.77	1.37	3.93
Embedded training, no B-W	0.28	0.30	0.83	1.41	4.24
Embedded training, Duration constraints	0.36	0.18	0.46	1.01	2.85

Table 2: Summary of Recognition Results

5.1 Isolated Training

5.1.1 Full left-right model structure

For this experiment, the transition probability matrix was initially full upper-triangular, i.e. transitions from any state to any subsequent state were permitted, and the starting point for re-estimation allowed entry or exit at any state (see figure 1).

Insertion errors were few and deletions more common, accounting for over 25% of the errors; this reflects the difference in speaking styles between the training data (deliberate and clearly enunciated) and the test data (including a higher proportion of coarticulation and elision, with each word being spoken more quickly). See table 4 for the detailed results.

5.1.2 Bakis model structure

For this experiment, the transition probability matrix was heavily restricted; transitions from one state to itself, the following state, and the next state after that were permitted,

but no others (see figure 2). Entry and exit were only permitted through the initial and final states respectively.

Very little difference in performance was observed between these two experiments; in practice, the full left-right model tended to reduce to the Bakis model during training. See table 5 for details.

5.2 Embedded Training

5.2.1 Full left-right model structure

Use of the full upper-triangular probability matrix was motivated by the need for flexibility to allow for durational compression of words in fluent speech. The results from the first speaker were so poor (see table 6) that the experiment was not continued; most of the errors were insertions of the digit "one", which could be compressed to such an extent that very little of the signal was still present. Comparison of the word boundaries discovered during recognition with the spectrograms of the phrases showed that in some cases the word boundaries were not being placed correctly.

The initial estimates of the state acoustic output vectors are probably not good, as they have been formed from isolated utterances; if the structure is tightly constrained, then the model may be forced to align with the relevant parts of the words, but if complete freedom is allowed it is likely that the model will fail to capture the features of the word correctly. This is one explanation for the poor recognition results obtained with this model structure; the full left-right transition matrix with unconstrained entry and exit points was clearly too flexible for the problem being addressed.

5.2.2 Bakis model structure

The more restricted structure was used to try and eliminate spurious insertion errors, since the model imposes a minimum word length equal to the integral part of $(N+2)/2$, where N is the number of states. Comparison of the results in table 7 with the results of the previous section shows the restriction to be justified. The results also show a significant improvement over those obtained using isolated word training, particularly in deletion and substitution error rates.

5.2.3 Outer-loop only re-estimation strategy

Interestingly, although deletion and substitution error rates increased when this training scheme was used, the number of insertion errors went down. The reason for this is not clear; what does emerge from these results is that the "inner loop" optimisation procedure has less effect on the recognition performance than the "outer loop" template extraction. (See table 8).

5.2.4 Durational constraints

Applying the word duration model described earlier resulted in a marked improvement in recognition performance (see table 9), halving the number of substitution errors. Surpisingly, the drop in insertions and deletions was less marked. However, inspection of the results obtained from native English speakers shows a pattern more consistent with the expected effects of this duration modelling strategy (see table 3); the numbers of insertion and deletion errors dropped sharply for these speakers.

5.3 Results for Native English Speakers

As it has been found that recognition error rates are generally higher for non-native than for native speakers [8], the results of the experiments using native English speakers are summarised in table 3.

ALGORITHM	% WORD ERRORS				% PHRASE ERRORS
	INSERTIONS	DELETIONS	SUBSTITUTIONS	TOTAL	
Isolated training, Full L-R	0.15	0.33	1.00	1.49	4.08
Isolated training, Bakis model	0.13	0.31	0.90	1.33	3.62
Embedded training, Bakis model	0.19	0.07	0.22	0.48	1.33
Embedded training, no B-W	0.22	0.04	0.26	0.52	1.55
Embedded training, Duration constraints	0.04	0	0.19	0.22	0.67

Table 3: Summary of recognition results for native English speakers.

6 Conclusions

The recognition algorithm investigated here has similarities to both the Sakoe and the Bridle algorithms; the chief advantage of the two-pass approach is that it enables the full likelihood at the word level to be determined, and is thus compatible with the Baum-Welch training algorithm. The separation of the two parts of the algorithm is useful as it facilitates the addition of extra constraints at the word level. The major difference between the approach discussed here and the Sakoe algorithm is that prior knowledge of the number of words in a string cannot easily be exploited; however, it is not necessary to put any limit on the possible length of a phrase, as required by the Sakoe algorithm. It is not clear that any of the dynamic programming algorithms for connected word recognition will perform significantly differently from each other, and an investigation of their relative merits is outside the scope of this memorandum.

It was obvious that training on isolated words did not produce models adequate for use in connected word recognition; whether the model was highly unconstrained, to allow for compression of the words in fluent speech, or had a tightly specified transition matrix, to discourage insertion errors, the results were disappointing. In fact, the model structure had very little effect on the recognition performance; the full left-right model, when applied to isolated words, reduced to the Bakis model during re-estimation in most cases.

The embedded training procedure, based on that described in [15], gave rise to a marked improvement in performance, so long as the Bakis model structure was used; if a full upper-triangular transition probability matrix was used, the performance became far worse than in the isolated-word training case.

It was found that if the Baum-Welch algorithm was performed after the segmentation procedure, the performance showed some improvement, but whether this improvement was sufficient to justify the increase in computer time required is questionable. Durational constraints seem to be of more importance in connected speech than isolated words; even the very simple "minimal word length" constraint imposed by the Bakis model structure gave an improvement in performance, and the more sophisticated model which assumed a Normal distribution of word lengths reduced the error rate still further, with very little increase in computational load.

References

[1] 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 62, pp 1035-1074, 1983.

[2] M. J. Russell & A. E. Cook, 'Experiments in speaker- dependent isolated digit recognition', Proc. Institute of Acoustics Autumn conference, Windermere, Vol. 8: Part 7, pp 293-298, 1986.

[3] J. S. Bridle, M. Brown & R. Chamberlain, 'An algorithm for connected word recognition', Proc. IEEE International Conference on Acoustics, Speech & Signal Processing, pp 899-902, 1982.

[4] H. Sakoe, 'Two-level DP-matching–a dynamic programming-based pattern matching algorithm for connected word recognition', IEEE Trans. Acoustics, Speech & Signal Processing, Vol. ASSP-27, no. 6, pp 588-595, December 1979.

[5] C. S. Myers & L. R. Rabiner, 'A level building dynamic time warping algorithm for connected word recognition', IEEE Trans. Acoustics, Speech & Signal Processing, Vol. ASSP-29, no. 2, pp 284-297, April 1981.

[6] S. Nakagawa & M. M. Jilan, 'Syllable-based connected spoken word recognition by two-pass O(n) DP matching and Hidden Markov Models', Proc. IEEE International Conference on Acoustics, Speech & Signal Processing, pp 1117-1120, 1986.

[7] B. Keck, 'An algorithm for the recognition of continuous speech with hidden Markov modelling', NTG-Fachber, Vol. 94, pp 97-102, 1986.

[8] R. K. Moore, 'Speech technology in a multilingual NATO environment', Proc. Military Speech Tech '87, pp 16-20, November 1987.

[9] R. S. Vonusa, J. T. Nelson & J. G. Parker, 'NATO AC/243 (panel III RSG10) Language Database', Proc. US NBS Workshop on Standardisation for Speech I/O Technology, PP 225-228, Gaithersburg 1982.

[10] L. R. Rabiner, J. G. Wilpon & B. H. Juang, 'A continuous training procedure for connected digit recognition', Proc. IEEE International Conference on Acoustics, Speech & Signal Processing, pp 1065-1068, April 1986.

[11] J. N. Holmes, 'The JSRU channel vocoder', Proc. IEE, 127, Pt.F(1), pp 53-60, 1980.

[12] T. H. Crystal & A. S. House, 'Characterisation and modelling of speech-segment durations', Proc. IEEE International Conference on Acoustics, Speech & Signal processing, pp 2791-2794, April 1986.

[13] M. J. Russell & A. E. Cook, 'Experimental evaluation of duration modelling techniques for automatic speech recognition',Proc. IEEE International Conference on Acoustics, Speech & Signal Processing, pp 2376-2379, April 1987.

[14] S. E. Levinson, 'Continuously variable duration hidden Markov models for automatic speech recognition', Computer Speech & Language, Vol. 1, No. 1, pp 29-45, 1986.

[15] L. R. Rabiner & S. E. Levinson, 'A speaker-independent, syntax-directed,connected word recognition system based on hidden Markov models and level building', IEEE Trans. Acoustics, Speech & Signal Processing, Vol. ASSP-33, no. 3, pp 561-573, June 1985.

Appendix A Results of Recognition Experiments

In the tables that follow, the speakers are identified by their initials; the other letters (A,B,C or D) denote the RSG10 list used for the recognition experiment. For example, the letters "MT - A" indicate speaker MT reading list 3A. Female speakers are denoted by (f).

ISOLATED TRAINING : FULL L-R MODEL					
SPEAKER FILE	NATIVE LANGUAGE	WORD ERRORS			PHRASE ERRORS
		INSERTIONS	DELETIONS	SUBSTITUTIONS	
BB - A	German	6	0	0	5
BB - D	German	5	1	1	7
FN - A	French (f)	4	1	3	8
FN - D	French (f)	6	4	11	18
GG - A	German	0	1	1	2
GG - B	German	0	0	2	2
GG - C	German	0	1	1	2
GG - D	German	0	1	2	3
GR - A	British (f)	1	1	1	3
GR - B	British (f)	1	0	1	1
GR - C	British (f)	3	0	2	5
GR - D	British (f)	0	0	0	0
HK - A	German	0	0	2	2
HK - D	German	0	0	0	0
HO - A	German (f)	0	0	0	0
HO - D	German (f)	1	0	0	1
JM - A	French	0	8	5	13
JM - B	French	0	6	11	15
JM - C	French	0	11	9	18
JM - D	French	0	11	10	18
JP - A	American	0	1	1	2
JP - D	American	0	0	2	2
KJ - A	American	0	0	0	0
KJ - B	American	0	0	0	0
KJ - C	American	0	0	0	0
KJ - D	American	0	0	0	0
LD - A	Dutch (f)	0	2	9	11
LD - D	Dutch (f)	0	0	22	21
LP - A	Dutch	0	4	12	14
LP - B	Dutch	0	3	9	10
LP - C	Dutch	3	4	6	13
LP - D	Dutch	1	5	7	11
MP - A	American (f)	0	0	1	1
MP - D	American (f)	0	0	0	0
MT - A	British	0	1	3	4
MT - B	British	0	0	2	2
MT - C	British	0	0	1	1
MT - D	British	0	0	6	6
MW - A	British	0	2	1	2
MW - B	British	0	0	3	3
MW - C	British	0	3	1	3
MW - D	British	0	2	4	5
RM - A	British	1	2	2	5
RM - B	British	0	1	4	4
RM - C	British	0	1	0	1
RM - D	British	0	0	3	2
SS - A	American	0	0	1	1
SS - D	American	0	0	0	0
TV - A	Dutch	0	2	5	7
TV - D	Dutch	0	4	11	14

Table 4: Results for Isolated Training with Full L-R Model

ISOLATED TRAINING : BAKIS MODEL					
SPEAKER FILE	NATIVE LANGUAGE	WORD ERRORS			PHRASE ERRORS
		INSERTIONS	DELETIONS	SUBSTITUTIONS	
BB - A	German	5	0	0	4
BB - D	German	5	1	1	6
FN - A	French (f)	4	0	3	7
FN - D	French (f)	6	4	11	18
GG - A	German	0	1	1	2
GG - B	German	1	0	2	3
GG - C	German	0	1	1	2
GG - D	German	0	1	3	3
GR - A	British (f)	1	1	1	3
GR - B	British (f)	1	0	1	1
GR - C	British (f)	2	0	2	4
GR - D	British (f)	0	0	1	1
HK - A	German	0	0	2	2
HK - D	German	0	0	0	0
HO - A	German (f)	0	0	0	0
HO - D	German (f)	1	0	0	1
JM - A	French	1	8	6	14
JM - B	French	0	6	10	14
JM - C	French	0	11	10	18
JM - D	French	0	11	10	18
JP - A	American	0	1	1	2
JP - D	American	0	0	3	3
KJ - A	American	0	0	0	0
KJ - B	American	0	0	0	0
KJ - C	American	0	0	0	0
KJ - D	American	0	0	0	0
LD - A	Dutch (f)	0	2	11	12
LD - D	Dutch (f)	0	0	20	19
LP - A	Dutch	0	5	8	13
LP - B	Dutch	0	3	7	8
LP - C	Dutch	1	4	5	10
LP - D	Dutch	1	6	9	13
MP - A	American (f)	0	0	1	1
MP - D	American (f)	0	0	0	0
MT - A	British	0	1	2	3
MT - B	British	0	0	3	3
MT - C	British	0	0	1	1
MT - D	British	0	0	3	3
MW - A	British	0	2	1	2
MW - B	British	0	0	3	3
MW - C	British	0	3	2	3
MW - D	British	0	1	3	4
RM - A	British	1	2	2	5
RM - B	British	0	1	3	3
RM - C	British	0	0	0	0
RM - D	British	0	0	1	1
SS - A	American	0	0	1	1
SS - D	American	0	0	0	0
TV - A	Dutch	0	2	5	7
TV - D	Dutch	0	3	10	12

Table 5: Results for Isolated Training with Bakis Model

EMBEDDED TRAINING : FULL L-R MODEL					
SPEAKER FILE	NATIVE LANGUAGE	WORD ERRORS			PHRASE ERRORS
		INSERTIONS	DELETIONS	SUBSTITUTIONS	
BB - D	German	148	5	1	50

Table 6: Results for Embedded Training with Full L-R Model

EMBEDDED TRAINING : BAKIS MODEL					
SPEAKER FILE	NATIVE LANGUAGE	WORD ERRORS			PHRASE ERRORS
		INSERTIONS	DELETIONS	SUBSTITUTIONS	
BB - D	German	5	1	0	6
FN - D	French (f)	1	0	1	2
GG - B	German	0	0	0	0
GG - C	German	0	0	0	0
GG - D	German	0	0	0	0
GR - B	British (f)	2	0	0	2
GR - C	British (f)	2	0	0	2
GR - D	British (f)	0	0	0	0
HK - D	German	0	0	0	0
HO - D	German (f)	0	0	0	0
JM - B	French	0	2	5	5
JM - C	French	2	1	5	8
JM - D	French	5	3	5	12
JP - D	American	1	0	0	1
KJ - B	American	0	0	0	0
KJ - C	American	0	0	0	0
KJ - D	American	0	0	0	0
LD - D	Dutch (f)	1	0	10	11
LP - B	Dutch	0	0	2	2
LP - C	Dutch	0	1	0	1
LP - D	Dutch	0	0	3	3
MP - D	American (f)	0	0	0	0
MT - B	British	0	0	0	0
MT - C	British	0	0	1	1
MT - D	British	0	0	2	2
MW - B	British	0	0	0	0
MW - C	British	0	2	1	3
MW - D	British	0	0	2	2
RM - B	British	0	0	0	0
RM - C	British	0	0	0	0
RM - D	British	0	0	0	0
SS - D	American	0	0	0	0
TV - D	Dutch	0	1	1	2

Table 7: Results for Embedded Training with Bakis Model

EMBEDDED TRAINING : NO B-W RE-ESTIMATION					
SPEAKER FILE	NATIVE LANGUAGE	WORD ERRORS			PHRASE ERRORS
		INSERTIONS	DELETIONS	SUBSTITUTIONS	
BB - D	German	5	1	0	6
FN - D	French (f)	1	0	1	2
GG - B	German	0	0	0	0
GG - C	German	0	0	0	0
GG - D	German	1	0	1	2
GR - B	British (f)	2	0	0	2
GR - C	British (f)	3	0	0	3
GR - D	British (f)	1	0	0	1
HK - D	German	0	0	0	0
HO - D	German (f)	0	0	0	0
JM - B	French	0	2	6	6
JM - C	French	1	3	5	9
JM - D	French	0	3	5	8
JP - D	American	0	0	0	0
KJ - B	American	0	0	0	0
KJ - C	American	0	0	0	0
KJ - D	American	0	0	0	0
LD - D	Dutch (f)	0	0	9	9
LP - B	Dutch	0	0	2	2
LP - C	Dutch	0	2	0	2
LP - D	Dutch	0	2	4	6
MP - D	American (f)	0	0	0	0
MT - B	British	0	0	1	1
MT - C	British	0	0	0	0
MT - D	British	0	0	1	1
MW - B	British	0	0	0	0
MW - C	British	0	1	1	2
MW - D	British	0	0	4	4
RM - B	British	0	0	0	0
RM - C	British	0	0	0	0
RM - D	British	0	0	0	0
SS - D	American	0	0	0	0
TV - D	Dutch	0	1	1	2

Table 8: Results for Embedded Training without Baum-Welch Re-estimation

EMBEDDED TRAINING : DURATIONAL CONSTRAINT					
SPEAKER FILE	NATIVE LANGUAGE	WORD ERRORS			PHRASE ERRORS
		INSERTIONS	DELETIONS	SUBSTITUTIONS	
BB - D	German	0	0	0	0
FN - D	French (f)	2	1	0	2
GG - B	German	0	0	0	0
GG - C	German	0	0	0	0
GG - D	German	0	0	0	0
GR - B	British (f)	0	0	0	0
GR - C	British (f)	2	0	0	2
GR - D	British (f)	0	0	0	0
HK - D	German	0	0	0	0
HO - D	German (f)	0	0	0	0
JM - B	French	2	1	4	6
JM - C	French	5	2	2	9
JM - D	French	5	4	1	9
JP - D	American	0	0	0	0
KJ - B	American	1	0	0	1
KJ - C	American	0	0	0	0
KJ - D	American	0	0	0	0
LD - D	Dutch (f)	1	0	5	6
LP - B	Dutch	0	0	3	3
LP - C	Dutch	0	0	0	0
LP - D	Dutch	0	0	3	3
MP - D	American (f)	0	0	0	0
MT - B	British	0	0	0	0
MT - C	British	0	0	1	1
MT - D	British	0	0	1	1
MW - B	British	0	0	0	0
MW - C	British	0	0	1	1
MW - D	British	0	0	0	0
RM - B	British	0	0	1	1
RM - C	British	0	0	0	0
RM - D	British	0	0	1	1
SS - D	American	0	0	0	0
TV - D	Dutch	0	1	0	1

Table 9: Results for Embedded Training with Overall Duration Constraint

DOCUMENT CONTROL SHEET

Overall security classification of sheet ..... UNCLASSIFIED .....

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the box concerned must be marked to indicate the classification eg (R) (C) or (S))

1. DRIC Reference (if known)	2. Originator's Reference	3. Agency Reference	4. Report Security Classification
	MEMO 4099		Unclassified

5. Originator's Code (if known)	6. Originator (Corporate Author) Name and Location
778400	RSRE, SAINT ANDREWS ROAD, MALVERN, WORCS WR14 3PS

5a. Sponsoring Agency's Code (if known)	6a. Sponsoring Agency (Contract Authority) Name and Location

7. Title
EXPERIMENTAL EVALUATION OF ALGORITHMS FOR CONNECTED SPEECH RECOGNITION

7a. Title in Foreign Language (in the case of translations)

7b. Presented at (for conference papers) Title, place and date of conference

8. Author 1 Surname, initials	9(a) Author 2	9(b) Authors 3,4...	10. Date	pp. ref.
COOK A E			1987.11	21

11. Contract Number	12. Period	13. Project	14. Other Reference

15. Distribution statement

Descriptors (or keywords)

continue on separate piece of paper

Abstract Current Automatic Speech Recognition devices attempt to solve the connected word recognition problem by assuming that an unknown phrase is the output of a sequence of statistical word-models. Typically, these models are constructed using examples of words spoken in isolation; however, the acoustic patterns corresponding to words as they occur in fluent speech are quite different from those representing the same words spoken in isolation, and so the use in speech recognisers of models based on isolated utterances severely limits the performance of such devices. A method of extracting training utterances from fluent speech and constructing Hidden Markov Models (HMMs) from these templates, known as *Embedded Training*, is investigated here, in conjunction with a two-level algorithm for connected word recognition. The effects on recognition performance of various HMM training procedures are discussed, and experimental results are presented.

END

DATE
FILMED

8 88