DYNAMIC HMM SELECTION FOR CONTINUOUS SPEECH RECOGNITION

T. Hain & P.C. Woodland

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.
{th223,pcw}@eng.cam.ac.uk

ABSTRACT

In this paper we propose a dynamic model selection technique based on hidden model sequences (HMS). HMS modelling assumes, that not only the actual state sequence is unknown, but also the model sequence given a particular sentence. This allows more than one model to be used for a particular phone in a certain context. The most appropriate model is determined locally rather than a priori globally by the acoustic probability of that model together with a probability that this model is produced in a particular phone (or model) context. Experiments on the Resource Management corpus show significant improvements in word error rate over phonetically model– and state–tied triphone hidden Markov models (HMMs). Initial results on the Switchboard corpus also show improvements on a much more difficult task.

1. INTRODUCTION

In HMM-based continuous speech recognition ideally each possible sentence would be modelled with a separate Markov model. Since the set of possible sentences is far too large to be able to train models for them, a particular sentence is split into words and further into a sequence of phones. Each phone (together with its phone context) usually is associated uniquely with a particular HMM. The concatenation of phone models results in the overall sentence model.

We propose a dynamic model selection technique, based on the use of hidden model sequences (HMS). HMS modelling assumes that not only the actual state sequence is unknown, but also the model sequence given a particular sentence. An additional stochastic level is added between the phonetic representation of a sentence and its realisation as a Markov chain. Multiple levels of modelling the transition from phonological units to their acoustic representation are assumed to be more realistic than a single level approach [3]. An advantage of a multi–level approach is that insertion and deletion effects present in spontaneous speech can be modelled on a lower level than the phone representation.

In this paper the use of HMS is restricted to *a priori* fixed alignment of phone to model mappings (1:1 mappings), which allow more than one model to be used for a particular phone in a certain context. The restricted scheme can be viewed as a soft version of model or state–tying or stochastic lexicon modelling [7]. The most appropriate model is determined locally rather than a priori globally [6] by the acoustic probability of that model together with a probability that this model is produced in a particular phone (or even model) context. State-of-the-art HMM models are built by concatenating states (left-to-right modelling), which makes it possible to extend HMS modelling to the state level. The HMS approach in general also allows the incorporation of additional acoustically based information in the selection process.

Section 2 outlines the theory of hidden model sequences. In Section 3 possible realisations of a model sequence model are discussed and the particular solution implemented for dynamic model selection is presented. Section 4 gives details of the experimental setup followed by presentation of experiments on the Resource Management (RM) (Sec. 4.1) and Switchboard (Sec. 4.2) corpora. The final section presents conclusions and an outlook to future work.

2. HIDDEN MODEL SEQUENCES

The overall goal in the standard training procedure of HMM-based speech recognisers is the maximisation of the likelihood of an observation sequence $\bar{O}$ given a sentence or word sequence $\bar{W}$. The usual approach uses a dictionary to uniquely map the word sequence $\bar{W}$ onto a phone sequence $\bar{R}$. This phone sequence further is translated into a HMM model sequence $\bar{M}$. Since the mapping from phone to model sequence is unique, given the set of model parameters λ, we can write

$$P(\bar{O}|\bar{S},\lambda) = P(\bar{O}|\bar{M},\lambda)$$

The mapping from the model sequence to the observation sequence is stochastic and determined by the HMM parameters. If the mapping from phone to model sequence is assumed to be non-deterministic as well, an appropriate model has to be found. The likelihood of the observation sequence given a sentence is found by summing over all possible model sequences $M \in \Omega(\bar{R})$ associated with a certain phone sequence, i.e.

$$P(\bar{O}|\bar{S},\lambda) = \sum_{M\in\Omega(\bar{R})} P(\bar{O},M|\bar{R},\lambda)$$

Using the chain rule two levels of modelling are visible:

$$P(\bar{O}, M|\bar{R}, \lambda) = P(\bar{O}|M, \bar{R}, \lambda)P(M|\bar{R}, \lambda)$$

The first level describes the relationship of observations to models whereas the next level gives the mapping between models and phonological units. Of course additional levels could be added by stochastic mapping of word to phone sequences or even sentence to word sequences.

Under the assumption that each level only depends on the level immediately above, the probability of the observation sequence simplifies to

$$P(\bar{O}|\bar{S}, \lambda) = \sum_{M \in \Omega(\bar{R})} P(M|\bar{R}, \lambda)P(\bar{O}|M, \lambda) \quad (1)$$

2.1. Parameter Estimation

The Estimation-Maximisation (EM) [2] algorithm is used to locally optimise the likelihood in (1) by global optimisation of the auxiliary function. With the reestimated parameter set $\hat{\lambda}$ the auxiliary function is represented by

$$Q(\hat{\lambda}|\lambda) = \sum_M \sum_q P(\bar{O}, q, M|\bar{S}, \lambda) \log P(\bar{O}, q, M|\bar{S}, \hat{\lambda}) \quad (2)$$

The inner sum is taken over all possible state sequences $q \in \Omega(M)$ given a particular model sequence M. Note that the pair (q, M) represents the hidden or unknown data. Using

$$P(\bar{O}, q, M|\bar{S}, \lambda) = P(\bar{O}, q|M, \lambda)P(M|\bar{R}, \lambda)$$

and further assuming independent parameter sets for the HMM ϕ and the model sequence model (MSM) θ we can define individual auxiliary functions for each level:

$$\begin{aligned} Q_{\mathrm{HMM}}(\hat{\phi}|M, \phi) &= \sum_q P(\bar{O}, q|M, \phi) \log P(\bar{O}, q|M, \hat{\phi}) \\ Q_{\mathrm{MSM}}(\hat{\theta}|M, \bar{R}, \theta) &= P(M|\bar{R}, \theta) \log P(M|\bar{R}, \hat{\theta}) \end{aligned}$$

The first function is the well known auxiliary function for standard HMMs. The modified overall auxiliary function (2) becomes

$$\begin{aligned} Q(\hat{\lambda}|\lambda) = & \sum_M Q_{\mathrm{HMM}}(\hat{\phi}|M, \phi)P(M|\theta) + \\ & \sum_M Q_{\mathrm{MSM}}(\hat{\theta}|M, \bar{R}, \theta)P(\bar{O}|M, \phi) \end{aligned} \quad (3)$$

Note that since we assumed independent parameter sets the first sum describes HMM parameter reestimation whereas the second sum gives the reestimation formula for the MSM parameters. Each of the reestimation formulae only depends on initial parameters of the other model. Thus the two terms can be maximised independently, given initial model sets for both models. The EM steps necessary are

1. Compute statistics $P(M|\theta)$and $P(\bar{O}|\phi)$.
2. Maximise $\sum_M Q_{\mathrm{HMM}}(\hat{\phi}|M, \phi)P(M|\theta)$ with respect to $\hat{\phi}$.
3. Maximise $\sum_M Q_{\mathrm{MSM}}(\hat{\theta}|M, \bar{R}, \theta)P(\bar{O}|M, \phi)$ with respect to $\hat{\theta}$.
4. Goto 1. using the new parameter estimates for both models.

In HMM training the Viterbi approximation is often used to simplify the sum over all possible state sequences by replacing it with the most likely state sequence q^*. If we take a similar approach in HMS modelling, the sum over all possible model sequences in (1) is simplified by taking the most likely model sequence:

$$M^* = \arg\max_M P(M|\bar{R}, \lambda)P(\bar{O}|M, \lambda) \quad (4)$$

Introducing the maximum approximation into the reestimation formula (3) becomes

$$\begin{aligned} Q(\lambda, \bar{\lambda}) \simeq & \; Q_{\mathrm{HMM}}(\hat{\phi}|M^*, \phi)P(M^*|\theta) + \\ & \; Q_{\mathrm{MSM}}(\hat{\theta}|M^*, \bar{R}, \theta)P(\bar{O}|M^*, \phi) \end{aligned}$$

This shows that not only both products can be maximised independently, but also the second term in both products denoting the influence of the MSM on the HMM and vice versa can be omitted, since they are constant during maximisation.

The EM procedure simplifies to first finding the best model sequence and updating both models by separate maximisation of $Q_{\mathrm{MSM}}(\hat{\theta}|M^*, \bar{R}, \theta)$ and $Q_{\mathrm{HMM}}(\hat{\phi}|M^*, \phi)$, for which standard Baum-Welch reestimation of HMM parameters can be used.

3. MODEL SEQUENCE MODELLING

A model sequence model maps an arbitrary phoneme sequence, defined by a word sequence and a particular dictionary onto an arbitrary model sequence. In its most general form, nothing is known about the particular models and how they relate to phonemes. Naturally the probability of the complete sequences will be split into a product of more local decisions based on shorter phone strings. In that sense a model sequence model has two degrees of freedom, the alignment between phone and model strings and the range of possible models or model strings for a given phone or phone string.

For dynamic model selection the alignment is assumed to be fixed, i.e. each phone can produce an *a priori* fixed number of consecutive models where each of the models may be chosen from a predetermined subset of all HMM models. Fixed alignment reflects the situation present in model or state-tied HMMs. Variable alignment, as for example used in [1], is capable of modelling insertions or deletions which most likely is relevant when dealing with spontaneous speech. In the case of fixed alignment a straightforward N-gram approach can be used.

$$P(M|\bar{R}) = \prod_{t=1}^{L} P(m_t|m_1...m_{t-1}, \bar{R})$$

If the history is restricted to the left and right phone neighbours and the dependency on the previous models is removed the HMS counterpart for triphone models is given by

$$P(M|\bar{R}) = \prod_{t=1}^{L} P(m_t|\bar{r}_{t-1}, \bar{r}_t, \bar{r}_{t+1}) \tag{5}$$

Each phone has a set of models associated with it which are assigned different probabilities depending on the left and right phone context. In recognition each model is a possible realisation of a particular phone. This allows a local decision on the most appropriate model rather than *a priori* selection by context and for example phonetic decision trees [6]. The approach can also be viewed as soft model tying. The most common implementation of phone models is concatenation of states ("pearls on a string"). Thus an extension to soft state tying is easily achieved by assuming that each of the states is an independent model. The single distribution for each phone context shown in (5) is replaced by a separate distribution for each state.

Since limited training data is available the model sequence model has to be able to deal with unseen phone contexts in addition to unseen models in a certain phone context. Both problems can be solved by the use of lower order statistics, which only takes for example the left context into account. For unseen models the distributions were smoothed using discounting. Methods commonly used in language modelling have been tested. A slightly modified version of Turing-Good discounting and absolute discounting have shown best performance, whereas Witten–Bell discounting [5] in general gave smoother, but slightly poorer results. More reliable smoothing was obtained by using the Katz–backoff scheme.

4. EXPERIMENTS

Experiments using the hidden model sequence paradigm and the assumption of fixed alignment have been conducted both on the Resource Management (RM) and on a subset of the Switchboard Corpus (SWBD). Whereas RM has been chosen to initially prove the viability of the approach, the effects present in spontaneous speech are the main target of this work.

Since the decoding networks are much larger than those for standard HMMs, results were obtained using N-best or lattice rescoring. RM results using the standard word–pair network decoding have also been obtained. To be able to build decoders with reasonable computational cost the MSM distributions had to be pruned to retain HMMs associated with 95-97% of the probability mass in each context.

All experiments have been initialised using model or state–tied HMMs, which also served to provide reference results for that particular experiment. The context was removed from each triphone model and the model was assumed to be a possible candidate for realisation of a certain phone. The number of possible candidates for different phones showed great variation on both corpora. Using the initial models a first alignment assuming uniform model distributions over all models in each phone context gave the first single best model sequence corresponding to (4). Using that model sequence the MSM N-grams were counted and HMM parameters reestimated. Naturally this very first step gave the largest improvement in log-likelihood. After initialisation in each reestimation iteration a new model sequence was found using the MSM and HMM parameters obtained in the previous iteration.

Although on RM experiments have been conducted by training from scratch rather than starting with a mixture density system, no improvements to the approach described above have so far been obtained.

4.1. Resource Management

The speaker independent Resource Management corpus consists of 3990 sentences of training data and 1200 sentences of test data split into the feb89, oct89, feb91 and sep92 evaluation sets. All systems in test use crossword triphones, a dictionary with one pronunciation per word and the standard RM word–pair grammar. For model-tied experiments the MSM used a modified version of Turing-Good discounting combined with Katz-Backoff to a bigram with left phone context and a further backoff to the unigram.

#mix	HMM	HMS-HMM	Δ% WER
1	7.38	6.68	-9.5
2	4.92	4.53	-7.9
4	3.79	3.24	-14.5

Table 1: %WER and relative improvement on the RM feb89 test set using 1153 crossword triphone HMMs and various number of mixture components (#mix). Both HMM and HMS-HMM use the same number of HMM parameters. HMS-HMM results were obtained by 20-best rescoring

First experiments using HMS at the model level with different number of mixture components and model-tied HMMs were conducted. Table 1 shows % word error rates (WER) for systems using 1153 triphone models. The largest gain in WER was obtained using 4 mixture components. Table 2 shows results on all test-sets. The overall result shows a significant improvement in WER of 14.9% relative.

System	feb89	oct89	feb91	sep92	total
HMM	3.79	5.51	4.91	8.28	5.63
HMS-HMM	3.16	5.10	3.86	6.99	4.79

Table 2: %WER on all RM test sets using a 1153 model crossword triphone models both for HMM and HMS-HMM. HMS-HMM results were obtained using 40-best rescoring.

The next experiment was a comparison with a state of the art state-tied triphone system. The reference HMM system is our best system using a single pronunciation dictionary on RM. The HMS-HMM was trained in a similar fashion as above. The results presented in Table 3 were obtained using network based decoding and the model distributions were pruned at a 95% level both in training and test. The overall relative reduction in WER was 16.9%. It is important to note that the performance depends on the scale factor used for the MSM. However WER varies smoothly with the MSM scale factor and identical scale factors for training and test appear to give the best results unless the scale factor is suboptimal.

System	feb89	oct89	feb91	sep92	total
HMM	3.16	3.80	3.30	6.17	4.11
HMS-HMM	2.77	3.13	2.62	5.20	3.43

Table 3: %WER on all RM test sets using crossword state-tied triphone models with a total of 1581 states for both HMM and HMS-HMM with decoding using a word-pair grammar.

4.2. Switchboard

Switchboard is probably one of the most difficult English corpora currently available. The highly spontaneous speech and the telephony bandwidth give word error rate results in the range between 30–50%. MSM-HMM models have been trained in a similar fashion as above, however now using an average of 1.12 pronunciation variants per word in a 24157 word dictionary. Initial experiments on this corpus used an 18 hour Switchboard subset for training (MiniTrain) [4].

A state-tied crossword triphone system built in the same way as those on RM showed poorer performance than the baseline on a half hour test-set (MTtest). Thus some changes were made to the original system. First of all the backoff model was changed to use a mixture of a right and left bigram context. Furthermore events seen only once in the training data have been ignored in MSM counting and absolute discounting showed slightly better performance. These improvements resulted in no gain on the first test-set but gave better results on a second half hour test set (WS96DevSub), a subset of the development test set used at the 1996 summer workshop at Johns Hopkins University. WER performance using both test sets shows an improvement of 0.77% absolute (see Table 4). However, error rates were poorer for only 8 out of 25 speakers.

System	MTtest	WS96DevSub	total
HMM	43.68	46.32	45.04
HMS-HMM	43.68	44.83	44.27

Table 4: Results on SWBD using crossword state-tied triphones models trained on 18 hours of speech and two test sets with each containing approximately half an hour of speech.

5. Conclusions

The proposed system using hidden model sequences was shown to give substantial improvement in performance compared to standard HMMs on Resource Management data. However smaller improvements have been achieved on the much more difficult Switchboard task. Further investigation will be necessary to find more appropriate discounting methods for the extremely sparse model distributions and better modelling of unseen contexts. An extension of this technique to variable length alignments is thought to be capable of capturing spontaneous speech effects present in the Switchboard corpus. Furthermore the proposed scheme is expandable to incorporate additional acoustic information in higher level models.

Acknowledgements

We would like to thank BBN for providing the MiniTrain training and MTtest test set definitions and JHU for the WS96Dev test set definition. This work was in part supported by GCHQ.

6. REFERENCES

1. S. Deligne and F. Bimbot, Inference of variable-length linguistic and acoustic units by multigrams, *Speech Communication* 23, pp. 223-241, 1997
2. A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Vol. 39 Ser.B, pp. 1-38, 1977
3. L. Deng, Speech recognition using the auto-segmental representation of phonological units with the interface to the trended HMM, *Speech Communication* 23, pp. 211-222, 1997
4. T. Hain, P.C. Woodland, T.R. Niesler and E.W.D. Whittaker, The 1998 HTK System for Transcription of Conversational Telephone Speech, *Proc. ICASSP'99*, Vol. 1, pp. 57-60, Phoenix
5. I.H. Witten and T.C. Bell, The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression", *IEEE Transactions on Information Theory*, Vol 37, No. 4, July 1991
6. S.J. Young, J.J. Odell and P.C. Woodland, Tree-Based State Tying for High Accuracy Acoustic Modelling, *Proc. DARPA HLT Workshop*, March 1994
7. S.-J. Yun and Y.-H. Oh, Stochastic Lexicon Modelling for Speech Recognition, *IEEE Signal Processing Letters*. Vol. 6, No. 2, pp. 28-30, February, 1999