MODELLING SUB-PHONE INSERTIONS AND DELETIONS IN 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

Recently, an extension to standard hidden Markov models for speech recognition called Hidden Model Sequence (HMS) modelling was introduced. In this approach the relationship between phones used in a pronunciation dictionary and the HMMs used to model these in context is assumed to be stochastic. One important feature of the HMS framework is the ability to handle arbitrary model to phone sequence alignments. In this paper we try to exploit that capability by using two different methods to model sub-phone insertions and deletions. Experiments on the Resource Management (RM) corpus and a subset of the Switchboard corpus show that, relative to standard HMM baseline, a reduction word error rate (WER) of 24.3% relative can be obtained on RM and 2.4% absolute on Switchboard.

1. Introduction

The dominant techniques in acoustic modelling for speech recognition are based on the use of Hidden Markov Models (HMMs). While there have been a number of advances in the use of HMMs in recent years, simple left-to-right chain topology models are still used in most recognition systems. Even though this "pearls-on-a-string" modelling approach is widely believed to be suboptimal, more complex schemes are rarely found in state-of-the-art systems.

In a standard HMM recognition system, each phone (in context) is uniquely associated with a particular HMM. Most commonly the associated HMM is found using phonetic decision trees which make use of local phonetic context [?]. The decision trees are usually built using an approximation of the models e.g. using a single Gaussian distribution per state when targeted at mixture Gaussian distributions. The decision tree technique allows particular model states (or complete models) to be used in a variety of phone contexts and the sets of shared states for each context to fill a predetermined model topology. The most common topology is a three state left-to-right model with only self transitions or transitions to the next state (i.e. no *skip* transitions).

Pronunciation variants or pronunciation networks are additional methods for model selection. Pronunciation modelling is considered to be particularly important for spontaneous speech, where the variability of pronunciation, stress and speaking rate is considerably greater than for read speech. Phone level transcriptions of spontaneous speech show significant differences between the canonical pronunciations in the dictionary and their actual realisation [?]. These differences can be described as substitutions, deletions and insertions where the latter appear to play a fairly minor role. However explicit modelling of these variations using pronunciation networks has brought only moderate improvements [?]. Furthermore, an analysis of spontaneous speech data in [?] showed that changes from canonical dictionary forms to surface forms can occur at the sub-phone level.

We have previously introduced the Hidden Model Sequence (HMS) modelling framework [?] in which the unique mapping from dictionary phones to HMMs was replaced by a stochastic mapping. This stochastic mapping uses an additional model sequence model (MSM) and in the HMS-HMM approach the model sequence for a particular pronunciation sequence is hidden. The most appropriate HMM to use is dependent on the particular acoustics rather than determined *a priori* using decision trees. The straight-forward HMS implementation used in [?] was constrained to use exactly one HMM for each dictionary phone[1] and was aimed at better handling within-phone variability. In this paper the use of HMS-HMMs which allow insertions and deletions of HMM states is presented and either a variation of the usual N-gram approach to MSMs or the general multigram method [?] is used to map from a dictionary phone sequence to an arbitrary length HMM sequence.

The rest of this paper is organised as follows. The next section briefly explains the framework and theory of HMS-HMMs and implementational issues. The following section discusses approaches for modelling of sub-phone insertions and deletions and explains two strategies that have been investigated. This is followed by an experimental evaluation of the techniques on the Resource Management and Switchboard corpora.

2. Hidden Model Sequence Modelling

The standard training procedure of HMMs for speech recognition maximises the likelihood of an acoustic feature sequence O given transcriptions of training data. In most cases a dictionary is used to translate the word sequence

[1] For state-level HMS-HMMs there were three HMMs per dictionary phone.

into a phone transcription sequence R which in turn is modelled by a (deterministic) sequence of context dependent HMMs. Therefore, knowledge of R is equivalent to knowing the HMM sequence M, or $p(O|R) = p(O|M)$. In contrast, HMS-HMMs assume a stochastic mapping:

$$p(O|R) = \sum_{M} p(O, M|R) = \sum_{M} p(O|M)P(M|R) \qquad (1)$$

where the additional stochastic layer $P(M|R)$ is the model sequence model (MSM) and the sum is taken overall all possible HMM sequences for the particular phone sequence. In (**??**) it is assumed that the probability of a stochastic sequence depends only on the sequence of the next level up, not on sequences higher up in the hierarchy. Note that no assumption about the length or alignment of the two sequences M and R has been made. The value of $p(O|M)$ is obtained using a standard HMM set. As shown in [**?**], when a model $P(M|R)$ is defined, the Expectation-Maximisation (E-M) algorithm can be used to locally maximise the overall likelihood $p(O|R)$. The Viterbi approximation at the model sequence level allows independent optimisation of the HMM and MSM parts and thus greatly simplifies the implementation. It has been used for all experiments in this paper.

If both sequences M and R are assumed to be aligned 1:1, the obvious realisation of an MSM is the N-gram model. A triphone model equivalent formulation (i.e. dependence on immediate left and right phonetic context) is given by

$$P(M|R) = \prod_{t=1}^{N} P(m_t|r_{t-1}r_t r_{t+1}) \qquad (2)$$

where N denotes the number of symbol pairs and m_t and r_t denote the individual models and phones respectively. For each phone a set of potential HMMs exists. All of these models in parallel constitute the new phone model, where the *model distribution* gives the probability of each model being used in a particular phone context. This scheme allows for modelling of phone or sub-phone substitutions.

To deal with unseen events in training, standard discounting schemes and Katz backoff to an interpolated left and right biphone and further to a context independent phone distribution are used. Witten-Bell discounting [**?**] was found to give superior performance over other schemes and has been used for all the HMS-HMM experiments reported here. HMS-HMMs can be implemented at a phone or sub-phone (state) level. The latter makes the model distributions as denoted in (**??**) dependent on the position within a phone.

In practice, a large number of parallel models is computationally too expensive. Thus the distributions over all models within a certain phone context are pruned to retain only 95% of the probability mass. In a similar way to how language models are used in recognition, the MSM scores are scaled in the log domain with an experimentally found scale factor. In this paper HMS-HMMs without insertion/deletion modelling are initialised from standard phonetic decision tree clustered HMMs and the set of possible models for a particular phone is derived by stripping the triphone context from the standard HMMs.

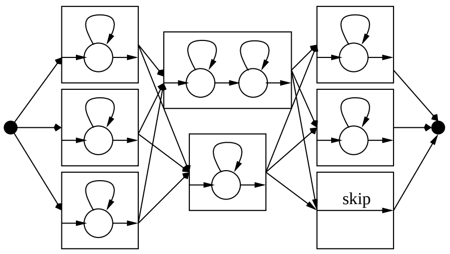

Figure 1: HMS-HMM topology on a state level with state insertions and deletions.

3. Modelling Sub-Phone Insertions and Deletions

Within the HMS framework several possibilities are available to model insertions and deletions at a sub-phone level. One option is to further assume fixed phone (in context) to model alignment. Since MSMs refer to HMMs in a purely symbolic way, the underlying topology of an HMM is arbitrary. The model may have multiple states or it even may contain no emitting state at all, in which case the model is referred to as "skip" model. Figure **??** shows an example of a HMS phone model with insertions and deletions. The objects in boxes are standard HMMs and the links in-between carry the context-dependent model distribution. If the alignment remains fixed, training and model construction schemes may be kept identical to the ones described in section **??**. However the models themselves cannot be drawn from an existing HMM set, as was done for substitution modelling and therefore alternative methods for initialisation are required.

3.1. Multigrams

Arbitrary phone to model mappings can be implemented using multigrams [**?**]. The theory of multigrams allows optimisation of the joint probability of two sequences. Theoretically an arbitrary length subsequence (string) in the phone sequence can produce an arbitrary length string in the model sequence. An empty string is not permitted and practically the string lengths have to be constrained. The set of possible model strings for a particular phone string is found automatically within an E-M training scheme, where the co-segmentation of the two sequences is used as a hidden parameter. In the original multigram work [**?**] the joint probability $P(M, R)$ served as the objective function for optimisation. MSMs however only require an estimate for $P(M|R)$, which only allows arbitrary length model strings and represents the case of 1:M mappings. In order to enable a certain degree of variation in sequence alignment the longest strings have to contain at least two models. Since multigrams are affected by data sparsity problems, the additional use of multiple phone strings seems to be infeasible.

Multigram training requires the optimisation of $P(M|R) = \sum_L P(M, L|R)$ where the hidden parameter L represents a particular segmentation of the model sequence (i.e. string boundaries) with exactly the same length as R. The E-M algorithm can be implemented efficiently using a forward/backward scheme:

$$\alpha(\tau,t) = P(M_1^\tau|R_1^t) \quad (3)$$
$$\beta(\tau,t) = P(M_{\tau+1}^{N_M}|R_{t+1}^{N_R}) \quad (4)$$

where τ and t are time indices and the notation M_1^τ denotes a subsequence of M including the boundary elements 1 and τ. The forward and backward variables can be computed using the recursions:

$$\alpha(\tau,t) = \sum_{k=1}^{K} \alpha(\tau-k,t-1)P(m_{\tau-k+1}^{\tau}|\mathbf{r}_t) \quad (5)$$
$$\beta(\tau,t) = \sum_{k=1}^{K} \beta(\tau+k,t+1)P(m_{\tau+1}^{\tau+k}|\mathbf{r}_{t+1}) \quad (6)$$

where K denotes the maximum number of models to be joined together and $\mathbf{r}_t$ stands for the phone plus context at position t, which in the triphone equivalent case is the triple (r_{t-1}, r_t, r_{t+1}). The reestimation of the probability of a particular model string μ of length k with a particular phone context ρ is

$$\hat{P}(\mu|\rho) = \frac{\sum_{t,\tau} \alpha(\tau-k,t-1)\beta(\tau,t)\delta_{\tau,t}^{\mu,\rho}}{\sum_{t,\tau} \alpha(\tau-k,t-1)\beta(\tau,t)\delta_t^{\rho}} \quad (7)$$

where $\delta_{\tau,t}^{\mu,\rho}$ and δ_t^ρ are indicator functions:

$$\delta_{\tau,t}^{\mu,\rho} = \begin{cases} 1 & M_{\tau-k+1}^{\tau} = \mu \text{ and } \mathbf{r}_t = \rho \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

$$\delta_t^\rho = \begin{cases} 1 & \mathbf{r}_t = \rho \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

3.2. Implementation

In practice the 1:1 phone to model mapping was relaxed gradually by a first introducing "skips" and then insertions in a second stage.

The first set of experiments (Scenario A) was aimed at using multigrams in a straight-forward fashion. For each phone context a large set of model strings was used. The model strings themselves consist of three single-state models. In a next stage the modelling of skips was implemented by adding all possible alternative models in which a single state was deleted. After several E-M steps using a standard N-gram MSM, in the second stage the N-gram model was replaced by a multigram model. Sharing of states between adjacent phone contexts was allowed.

Initial model strings for Scenario A were obtained from a state-level HMS-HMM system. All state-model triplets within each triphone context were collected and added to a new HMS-HMM set. The HMS-HMM then operates on the phone level. Since the number of such triplets is large, sparsity problems require a considerable amount of pruning of model distributions. The sparsity of distributions is the major disadvantage throughout all stages of this scenario, which even required hard limits to be placed on the number of allowable alternative models during training.

To overcome the above difficulties, a second approach, Scenario B, tried to make estimates for model probabilities more robust by using MSMs on a state level. An explicit skip HMM (i.e. with no output distribution) was added to the list of possible models within each phone context. A common initial value for the probability of these skip models in all contexts was chosen. After several E-M iterations an estimate of the probability of skips in each phone context was obtained. Two different second stages were tested. In one case all pairs of models from the same phone position have been added to the model distribution. This can be sufficiently modelled using N-grams. In the second case models from adjacent phone positions but within a particular phone were added. After some reestimation steps using N-grams, a multigram MSM with a maximum of two elements per model string was used.

4. Experiments

All insertion/deletion experiments used an HMS-HMM model set for initialisation, which itself was initialised from the baseline HMM model set (see [?] for details). Thus all model sets for a particular corpus have exactly the same number of states and mixture components per state.

Experiments have been conducted using both Resource Management (RM) and Switchboard. RM was chosen because of the small amount of training data and the single pronunciation dictionary in our standard HMM setup. The main focus of this work, however, lies in modelling of spontaneous speech. Experiments on the Switchboard corpus used a 18 hour training set of Switchboard-I (Swbd-I) data (Minitrain) and tests were conducted on two 30 minute Switchboard test sets (MTtest and WS96DevSub). The HMM baseline system has 2954 states and 12 mixture components and uses conversation side based cepstral mean and variance normalisation.

	MTtest	WS96DevSub	Overall
baseline HMM	43.68	46.32	45.04
baseline HMS	42.88	44.70	43.82
HMS model level	43.69	45.25	44.49
+ max model	43.36	44.63	44.01
+ deletions	43.41	44.38	43.91
+ multigram	43.14	44.17	43.67

Table 1: Scenario A: %WER on Switchboard using various model level HMS-HMMs. Results were obtained by rescoring trigram lattices.

Due to the problems of data sparsity, experiments for Scenario A have only been conducted on Switchboard. Table **??** shows word error rates (WERs) on both test sets for the baseline HMM and HMS-HMM systems and for the various stages within Scenario A. As can be seen the step towards phone unit modelling brings a performance degradation which can be partly recovered by setting a hard limit to the maximum number of models per phone context (max model). The addition of deletion modelling brings only minor improvements and still has a poorer WER than the HMS-HMM baseline. The use of multigrams finally brings slightly better performance than the baseline HMS-HMM system.

Scenario B was tested on both RM and Switchboard. 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. Our standard RM model sets have 1581 states and 6 mixture components. The dictionary contains 991 words with single pronunciations. A word-pair grammar is used for recognition. Table **??** shows word error rates on all four test sets.

The baseline HMS-HMM system already gave a considerable improvement over the standard HMM models. A further slight improvement could be made by added skip models. However on this data the gain from added insertions was larger. Modelling of sub-phone insertions and deletions within the same phone position gave a 5.7% relative reduction in word error rate over the HMS baseline and thus a 24.3% relative reduction over the HMM baseline. Using skips brought a 0.09% absolute reduction both in word deletions and substitutions but an increase in insertions. Modelling insertions further reduced the number of deletions while other substitution and insertion errors remained virtually unchanged. Using models from adjacent phone positions gave an overall word error rate of 3.20% and poorer results if further extended to using multigrams. A potential reason for this are overestimated skip probabilities.

System	feb89	oct89	feb91	sep92	overall
HMM	3.16	3.80	3.30	6.17	4.11
HMS-HMM	2.62	3.20	2.54	4.81	3.30
+ skips	2.66	3.09	2.86	4.34	3.24
+ insertions	2.23	2.94	2.74	4.53	3.11

Table 2: Scenario B: %WER on RM using state level HMS-HMMs. Results have been obtained using Viterbi decoding with a single pronunciation dictionary and a word-pair grammar.

Table **??** shows Scenario B results on Switchboard. The difference in the HMS baseline (compared to Table **??**) is due to the silence modelling now used which is identical to that in the HMM baseline. Again the largest improvement of 1.6% WER absolute comes from straight-forward HMS modelling. However, in contrast to the results on RM a further WER reduction by 0.8% can be achieved by the use of skip models. Whereas the number of word deletions remained approximately the same, most of the improvement stems from a reduced number of word substitutions.

Again in contrast to RM, insertions brought no significant reduction in word error rate, even though a considerable increase in acoustic log-likelihood was observed. The probable reason for this is the broadness of the MSM distributions which could not be controlled even with stricter pruning. Thus an overall improvement of 2.4% WER absolute over the HMM constitutes the best result so far. Similarly to the situation on RM, the use of models from adjacent positions and multigrams actually gave considerably poorer performance.

System	MTtest	WS96DevSub	Overall
HMM	43.68	46.32	45.04
HMS-HMM	42.80	43.92	43.38
+skips	42.14	43.11	42.64
+insertions	41.81	43.29	42.57

Table 3: Scenario B: %WER on Switchboard obtained by rescoring of trigram lattices.

5. Conclusions

Sub-phone insertions and deletions have been modelled within the HMS framework. Although substantial improvements can be gained using HMS-HMMs in a straightforward fashion, only context-specific deletion modelling seems to give consistent performance improvements. One of the major reasons for this are the very broad model distributions obtained for insertions especially on Switchboard data which add confusion to the overall system. Future work in this area will have to concentrate on mechanisms which enforce compact distributions.

6. Acknowledgement

This work was in part supported by GCHQ.

7. REFERENCES

1. W. Byrne, M.Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraçlar, C. Wooters & G. Zavaliagkos (1998). Pronunciation modelling using a hand-labelled corpus for conversational speech recognition. *Proc. of ICASSP'98*, pp. 313-316, Seattle.
2. S. Deligne & F. Bimbot (1997). Inference of variable-length linguistic and acoustic units by multigrams. *Speech Communication*, Vol. 23, pp. 223-241.
3. S. Greenberg (1996) The Switchboard Transcription Project. 1996 LVCSR Summer Workshop Technical Reports. `http://www.icsi.berkeley.edu/real/stp`.
4. T. Hain & P. C. Woodland (1999). Dynamic HMM selection for continuous speech recognition. *Proc. EUROSPEECH'99*, pp. 1327-1330, Budapest.
5. M. Saraçlar & S. Khudanpur (2000). Pronunciation ambiguity vs pronunciation variability in speech recognition. *Proc. ICASSP'2000*, pp. 1679-1682, Istanbul.
6. I. H. Witten & T. C. Bell (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Trans. Information Theory*, Vol. 37, pp. 1085-1094.
7. S. J. Young, J. J. Odell & P. C. Woodland (1994). Tree-Based State Tying for High Accuracy Acoustic Modelling. *Proc. ARPA Human Language Technology Workshop*, pp. 307-312. Morgan Kaufmann.