THE HTK TIED-STATE CONTINUOUS SPEECH RECOGNISER

P.C. Woodland & S.J. Young

Cambridge University Engineering Department, England

ABSTRACT

HTK is a portable software toolkit for developing systems using continuous density hidden Markov models developed by the Cambridge University Speech Group. This paper describes speech recognition experiments using HTK based systems for the DARPA Resource Management (RM) task. In particular good performance is obtained using a tied-state triphone based multiple mixture approach. This system was used in the final DARPA RM evaluation (September 1992) and was found to perform at a similar level to the main DARPA systems, and yet be efficient in terms of the total of parameters and computational load. The results for that system are given along with some recent experiments that investigated the use of male-female modelling in a tied-state HMM system.

Keywords: Hidden Markov Models, Resource Management, State Clustering, HTK.

1. INTRODUCTION

This paper describes the use of HTK (HMM toolkit) in building speech recognisers for the DARPA Resource Management (RM) task. HTK is a software toolkit for building and manipulating systems that use continuous density hidden Markov models that has been developed by the Speech Group at Cambridge University Engineering Department over the last four and a half years. HTK is designed to be flexible enough to support both research and development of HMM systems and also to provide a platform for benchmark evaluations. It can be used to perform a wide range of tasks, including isolated or connected speech recognition using models based on whole word or sub-word units. HTK includes a software library as well as a number of tools (programs) that performs tasks such as coding data, various styles of HMM training including embedded Baum-Welch re-estimation, Viterbi decoding, results analysis as well as a flexible HMM editor. The current version of HTK, V1.4, is now in use at many sites worldwide. Full details of HTK V1.4 are given in [7] and an overview is given in [6].

A number of features of HTK make it especially suitable for performing large vocabulary continuous speech recognition on a corpus such as DARPA RM. In particular, HTK has a unique generalised parameter tying (sharing) mechanism that allows HMM systems to be constructed that are balanced between acoustic modelling detail (model complexity) and parameter estimation accuracy for a given training corpus. In practice, for any reasonably complex task, data will be limited for at least some models and either parameter tying or parameter smoothing will be necessary to obtain good performance. A large number of different styles of parameter sharing are possible using HTK . In particular it has been found that a triphone based system using multiple mixture output densities using state sharing between the corresponding states of some allophones of each phone gave good performance on the RM task. The tied states are chosen on the basis of acoustic similarity with the constraint that all states must have enough training data to estimate the parameters of the mixture distributions. This recogniser was included in the final September 1992 DARPA RM evaluation [6]. It is unique in that it was built using a publicly available 'off-the-shelf' package; it is relatively compact having about 800,000 parameters; it has reasonable decode times of about 20 seconds per sentence on a standard workstation; and it delivers recognition performance comparable to the current DARPA systems.

This paper first describes some of the experiments performed when building the tied-state system used for the September 1992 RM evaluation, and then some experiments using male/female modelling in the context of a tied-state continuous density HMM system.

2. SEP.'92 RM SYSTEM & EXPERIMENTS

This section describes experiments performed using HTK to build systems for the final DARPA RM evaluation (September 1992). Firstly some simple monophone systems (single mixture and multiple mixture) were built and then the tied-state multiple mixture triphone system were trained. The experiments described below were all performed using the standard tools in HTK V1.4. In addition some additional tools were written to perform tasks such as recognition network file creation, label file creation and dictionary manipulation. Furthermore a number of scripts were used for tasks such as data creation and running experiments.[1]

For all experiments the standard RM SI-109 training set was used (3990 sentences), and systems were evaluated using the speaker independent Feb'89, Oct'89, Feb'91 and Sep'92 test sets. Tests using both "no-grammar" (perplexity 991) and the standard RM word-pair grammar (perplexity 60) were run. All results were scored using the the standard NIST scoring software, although exactly the same results can be obtained using the standard HTK scoring software.

[1] These additional tools, scripts and instructions for building RM systems will be bundled with a future release of HTK.

Since the RM data is only labelled at the word level and no pronunciation information is given a pronunciation dictionary and corresponding phone set is needed. For the experiments reported here, the dictionary and phone set produced by CMU and listed in [3] were used. The phone set consists of 48 symbols (including silence) and the dictionary gives only a single pronunciation for each of the 991 lexical items.

The basic acoustic analysis used an observation vector consisting of 12 Mel frequency cepstral coefficients and log energy. To this the first differential of each of these components was appended to make a 26 dimensional observation vector. Some initial experiments showed that performance improved if the second differentials were also used increasing the dimensionality to 39. Perhaps somewhat surprisingly, these experiments showed that adding the second differential coefficients improved word-pair grammar word-error rates by between 2 to 3 % (absolute) over a wide range of system types from single mixture monophone HMMs to multiple mixture triphone models. Clearly more accurate systems are better able to make use of this detailed information and improve by a larger (relative) factor. Hence all of the experimental results reported here use the 39 dimensional parameter vectors which include static, first and second differential information.

Initially 48 models each with a three state left-to-right topology (mostly with no skip transitions), and single mixture diagonal covariance output distributions were defined. The models were initialised from previously trained TIMIT HMMs taking account of differences between the TIMIT and RM phone sets.[2] All subsequent training of the HMMs was performed using embedded Baum-Welch re-estimation. The single mixture monophone RM HMMs were trained and gave a word-error rate of 25.4% on the Feb'89 test-set with the word-pair grammar. This basic setup was then modified to allow an optional inter-word single-state silence model in both HMM training and during recognition. This decreased the error rate to 23.1%. Finally, the 32 commonest function words in the RM training data were modelled separately by using function-word specific phones. This increased the total model count (including the inter-word silence) to 130. A single mixture system was then trained for the 130 model set and the error rate again decreased to 17.3%.

Clearly it is beneficial to use both inter-word silence modelling and function-word specific phones and these were retained for building the more complex systems for the Sep'92 evaluation. However, for good performance it is clear that the single mixture models do not provide sufficient detail and that multiple mixture output distributions and/or context-dependent models are required. To calibrate the use of triphone based multiple mixture models, some monophone multiple mixture systems were first trained and tested, and then the tied-state multiple-mixture triphones were evaluated.

2.1. Multiple Mixture Monophone Systems

To calibrate the use of context-dependent multiple mixture models, firstly multiple-mixture context independent models (130 models) were built. If monophones are viewed in terms of triphone units then a monophone system can be thought of as tying all (context-dependent) models for the same phone together for all left and right contexts. In this case, there is sufficient data to train multiple mixture models for each of the phones. Furthermore, a relatively large number of mixtures will indeed be needed to accurately capture the variability of the data since a large number of contexts have been merged.

The 130 model multiple mixture system was trained starting from the single mixture monophone baseline by a iteratively using *mixture-splitting* [8] and then re-training. In this way two mixture, three mixture, five mixture, seven mixture, ten mixture and finally fifteen mixture systems were trained. The 15 mixture system contains about 463,000 parameters.

It was found that the addition of multiple mixtures caused a large decrease in word errors. For instance on the Feb'89 test data using the word-pair grammar a 2 mixture system reduced the word error rate to 11.3%; with three mixtures it was 9.0%; with five mixtures 7.8%; 10 mixtures 6.4% and finally the 15 mixture system had a word error rate of 5.7%. It should perhaps be noted that the error rate of this 15 mixture monophone system is lower than the error-rate of the best system reported at February 1989 DARPA Speech and Natural Language Workshop [4]. That system was considerably more complex and used cross-word triphones models, but did not model second differential data.

2.2. Tied-State Triphone System

It has been widely accepted that the sub-word units in a large vocabulary speech recognition system must be context-dependent for good performance (e.g. [5],[3]), and systems have been developed that use either word-internal triphones (i.e. models that are not dependent on context across a word boundary) or cross-word triphones. However for a triphone system, the data insufficiency problem is acute. For example, one word-internal triphone doesn't occur at all in the training data, and some 589 word-internal triphones have 10 or less occurrences. If any HMM based recogniser is to be successful, the system must be constructed so that each parameter is (fairly) well-trained. A number of methods of tying could be used, however, it is necessary to retain output distributions with sufficient acoustic detail whilst avoiding undertraining. The approach investigated here is to construct a tied-state HMM system.

State distributions for corresponding states in allophones of the same phone are clustered and tied if they are acoustically similar ensuring that there is enough training data for a constant number of mixture components across all distributions. This approach could be regarded as data-driven sub-phone modelling and, in that respect, is related to the work in [1].

The HTK tied-state multiple mixture triphone system was trained in a number of stages. Firstly 2427 non-tied single mixture word-internal triphones (and function word specific phones) were estimated. These models were initialised by cloning the the 130 single mixture monophone model set and re-training. This system gave 10.0% word errors on the Feb'89 test set using the word-pair grammar.

The corresponding states of each of the context-dependent allophones were then clustered using a hier-

[2] Equivalent results can be obtained with HTK using a "flat-start" i.e. all model output distributions are initialised to the same values.

archical furthest neighbour clustering alorgithm. This operates as follows. Each state starts as a single cluster and then the algorithm repeatedly merges pairs of clusters. Clusters to merge are chosen such that the newly formed cluster will have the minimum intra-cluster distance possible at that stage of the clustering process. The final number of clusters is determined by ensuring that the maximum intra-cluster distance is below a threshold. For these calculations, the intra-cluster distance is defined as the maximum distance between any pair of states within a cluster. In these experiments, the distance between the (single-mixture) distributions for each state is computed as the Euclidean distance between the means weighted by the geometric mean of the variances. An alternative distance measure that can be used in the clustering procedure is the divergence between the two distributions. This was used in the experimental evaluation of state clustering and state-tying reported in [8].

After this clustering stage, the distributions for sets of states are based on their acoustic similarity without accounting for the amount of training data available. Therefore, as a second stage, an iterative procedure is applied whereby any clusters for which the sum of the occupation counts for all states within a cluster (determined during the final re-estimation iteration) falls below a threshold are merged with the nearest cluster. This stage ensures that there will be enough data to train the distribution associated with the set of clustered states. Finally, the distributions associated with these clustered states are tied, as are the transition matrices across all allophones of each phone (130 transition matrices in total). The parameters of the HMMs were then re-estimated.

Applying this state clustering and tying procedure to the single mixture triphone system reduced the total number of distributions by 77% (7929 to 1811). However the total number of models with different sets of distributions was reduced by only 16% (2427 to 2029). This implies that a much more efficient clustering is possible at the state level than at the model level, which is how generalised triphones are often produced [3]. It is also interesting to note that often the clustering process automatically grouped allophones with contexts in the same broad phonetic class. After retraining, the single mixture state-clustered system gave a word-error rate of 7.7% on the Feb'89 test set with a word-pair grammar, i.e. 23% less errors than the untied single mixture triphone set with less than a quarter of the parameters. It is clear that the state-clustering procedure is highly effective in reducing the under-training problem.

Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	96.0	2.6	1.4	0.4	4.5	26.0
Oct'89	95.4	3.1	1.5	0.5	5.1	25.7
Feb'91	96.6	2.3	1.1	0.6	4.0	20.3
Sep'92	93.6	4.4	2.0	1.0	7.4	36.0

Table 1. % recognition results for September 1992 6 mixture tied-state triphone system (word-pair grammar)

To improve the acoustic detail of these state-clustered models, multiple mixture models were again created by using mixture-splitting and re-training. This process can now be performed without severe under-training problems for all distributions because the state clustering procedure ensures that there is sufficient training data for multiple mixtures. Starting with the single mixture state-clustered system, 2 mixture, 3 mixture, 4 mixture, 5 mixture and finally a 6 mixture system was trained containing about 857,000 parameters.

Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	83.8	12.5	3.7	1.5	17.7	69.0
Oct'89	81.7	14.0	4.3	2.0	20.3	77.3
Feb'91	84.8	12.7	2.5	1.9	17.1	65.7
Sep'92	78.3	16.8	4.9	2.8	24.5	78.0

Table 2. % recognition results for September 1992 6 mixture tied-state triphone system (no grammar)

As more mixtures are added to the tied-state system the recognition results steadily improve. The 2 mixture system has a word error rate of 5.7% with the word-pair grammar on the Feb'89 test set; with 3 mixtures it is 5.3%; 4 mixtures 5.0%; 5 mixtures 4.8% and 6 mixtures 4.5%— the six mixture tied-state system has 42% fewer errors than the single mixture tied-state system. The 2 mixture system performs as well as the 15 mixture monophone system, but contains fewer parameters and runs considerably faster at recognition time. The full results for the 6 mixture tied-state triphone system for all four speaker independent RM test sets are shown in Table 1 for word-pair grammar and in Table 2 for no-grammar testing.

The computational cost of the state-clustered system was measured. Using a HP 9000/730 computer with fairly light pruning enabled, it was found that in training the single mixture system required about half a second per utterance per iteration and the six mixture system just over two seconds per utterance per iteration. During recognition with a word-pair grammar the single mixture system required about 10 seconds per utterance, while the six mixture system took about 20 seconds per utterance.

The level of performance achieved by the six-mixture tied state system is comparable with that achieved by systems developed by the main DARPA sites, however it used neither male/female modelling nor cross-word triphones.

3. GENDER SPECIFIC MODELLING

To try and improve the performance of the tied-state HMM system, some further experiments have been performed in which separate HMMs have been constructed for the male and female speakers. A number of sites have shown that this approach can lead to improved performance, although less training data is available for each model parameter.

In the SI-109 RM training corpus, there are 31 female speakers (1160 utterances) and 78 male speakers (2830 utterances). For this comparison, since a rather smaller amount of data is available for the female speakers the systems were built without function-word specific models. Also these new sets of models were tested with the addition of intra-word optional silences in acronyms and hyphenated words.

Firstly, a new speaker independent tied-state system was built without function-word specific mod-

els (2369 models) and had 1655 states after state-clustering (from 7111 before clustering). The number of mixtures was increased to 6, and the results for this system when tested with the word-pair grammar are shown in Table 3. It will be noted from these results that the performance is better than that given in Table 1 for the Feb'89, Oct'89 and Feb'91 test sets, but worse for the Sep'92 set.

Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	96.2	2.4	1.4	0.3	4.1	25.7
Oct'89	95.7	2.8	1.5	0.6	4.8	25.3
Feb'91	96.7	2.2	1.1	0.5	3.8	22.0
Sep'92	92.9	4.8	2.3	0.9	8.0	37.3

Table 3. % recognition results for speaker independent 6 mixture tied-state system without function word specific phones (word-pair grammar)

A system was then built with a set of female models trained on the female data only, and the male models trained on the male training data (MF system). Sex independent silence models were used. After state-clustering there were 1048 female states and 1564 male states (from 7107 each). It will be noticed that the state-clustering algorithm automatically assigns fewer states to the female models since there is less female training data. Since there are rather more states in total in the male/female case than in the speaker independent case, 5 mixture models were trained.

Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	96.4	2.3	1.3	0.5	4.1	25.7
Oct'89	94.8	3.7	1.5	0.9	6.1	31.7
Feb'91	96.1	2.8	1.1	0.9	4.8	25.3
Sep'92	93.0	4.7	2.3	0.9	7.8	36.7

Table 4. % recognition results for MF 5 mixture tied-state system

The recognition network for the MF system consisted of two complete separate word-pair grammar networks in parallel. It was found that a slightly higher pruning threshold than usual was required to avoid occasionally pruning out the correct branch of the network, and the total recognition time was about 30% more than that required for a speaker independent system. The results for this configuration are given in Table 4. It can be seen that for all test-sets apart from the Sep'92 set there is no improvement in performance. In fact, on average, performance is about 10% worse than for the speaker independent case. However, performance improves for some speakers and degrades for others.

It was therefore decided to use the 5 mixture male/female models in parallel with the 6 mixture speaker independent models in a three branch network (MFI system). It was hoped that for the test speakers that were modelled accurately using the male/female models that these models would be used, otherwise the speaker independent branch would be preferred. The results for the MFI system are shown in Table 5. It can be seen that overall there is approximately a 3% decrease in word-error rate over the use of speaker independent models alone and a 12% reduction in error rate from using the male/female models alone.

Test Set	Words Corr	Subs Err	Del Err	Ins Err	Word Err	Sent Err
Feb'89	96.7	2.1	1.2	0.4	3.7	24.3
Oct'89	95.5	3.1	1.3	0.6	5.0	26.7
Feb'91	96.7	2.3	1.0	0.8	4.1	23.3
Sep'92	93.5	4.3	2.2	0.7	7.2	35.3

Table 5. % recognition results for MFI system

These results show that separate male/female and speaker independent models can be tested in parallel in this manner but give only rather small gains in performance. It may well be the case, however, that a larger improvement would be obtained by using MAP estimation of the male/female model parameters with speaker independent priors [2].

4. CONCLUSION

This paper has described the use of HTK for the DARPA Resource Management task, and in particular systems using state level tying. This multiple mixture tied state triphone system gave good performance. It is efficient in the way it uses parameters and has a relatively low computational cost. The error rates for this approach are considered to be very respectable considering that cross-word triphone models are not used. Additional experiments have shown that a further small reduction in error rate can be obtained by using gender-specific models if the speaker independent models are also used in parallel. Since HTK V1.4 is commercially available, it is straightforward to both duplicate and extend the systems described in this paper.

REFERENCES

[1] Hwang M-Y. & Huang X. (1992). Subphonetic Modelling with Markov States. *Proc. ICASSP'92*, Vol. I, pp. 33-36, San Francisco.

[2] Gauvain J-L. & Lee C-H. (1992). Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities. *Speech Communication*, **11**, 205-213.

[3] Lee K-F. (1989). Automatic Speech Recognition: The Development of the SPHINX System. Kluwer Academic Publishers, Boston.

[4] Pallett D.S. (1989). Speech Results on Resource Management Task. *Proc. DARPA Speech & Natural Language Workshop February 1989*, pp. 18-24, Philadelphia.

[5] Schwartz R., Chow Y., Kimball O., Roucos S., Krasner M. & Makhoul R. (1985). Context-Dependent Modelling for Acoustic-Phonetic Recognition of Continuous Speech. *Proc. ICASSP'85*, pp. 1205-1208, Tampa.

[6] Woodland P.C. & Young S.J. (1992) Benchmark DARPA RM Results with the HTK Portable HMM Toolkit. *Proc. DARPA Continuous Speech Recognition Workshop*, September 1992, Stanford.

[7] Young, S.J. (1992). HTK Version 1.4: User, Reference & Programmer Manual. Cambridge University Engineering Department. August 1992.

[8] Young S.J. & Woodland P.C. (1993) The Use of State Tying in Continuous Speech Recognition. *Proc. Eurospeech*, Berlin.