THE 1994 HTK LARGE VOCABULARY SPEECH RECOGNITION SYSTEM

P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev & S.J. Young

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, England.

ABSTRACT

This paper describes recent work on the HTK large vocabulary speech recognition system. The system uses tied-state cross-word context-dependent mixture Gaussian HMMs and a dynamic network decoder that can operate in a single pass. In the last year the decoder has been extended to produce word lattices to allow flexible and efficient system development, as well as multi-pass operation for use with computationally expensive acoustic and/or language models. The system vocabulary can now be up to 65k words, the final acoustic models have been extended to be sensitive to more acoustic context (quinphones), a 4-gram language model has been used and unsupervised incremental speaker adaptation incorporated. The resulting system gave the lowest error rates on both the H1-P0 and H1-C1 hub tasks in the November 1994 ARPA CSR evaluation.

1. INTRODUCTION

This paper describes recent improvements to the HTK large vocabulary speech recognition system. The system uses state-clustered mixture Gaussian cross-word triphone HMMs to allow an appropriate balance of acoustic modelling detail (model complexity) and parameter estimation accuracy for a given training corpus. The system decoder is able to integrate cross-word context-dependent acoustic models and N-gram language models (LMs) into a single recognition pass. The system was evaluated using the Wall Street Journal (WSJ) corpus and found to give very good performance [5].

Recently, our focus has been to optimise and further enhance the capabilities of the system. However experimentation using our original system was computationally very costly. Therefore the decoder has been extended so that it can produce a network of recognition alternatives stored in the form of a *word lattice* for each sentence. The information stored in these word lattices can be used for many purposes including the optimisation of system parameters, or investigating the use of alternative acoustic and language models.

This paper first gives an overview of the HTK large vocabulary recognition system and then the process of generating word lattices is described. The further development of the system for the November 1994 ARPA continuous speech recognition (CSR) evaluation is then detailed, which includes the use of refined acoustic models, a 65k word vocabulary, a 4-gram language model and the integration of unsupervised incremental speaker adaptation. The results show that the resulting system continues to give state-of-the-art performance.

2. SYSTEM OVERVIEW

This section gives an overview of the HTK large vocabulary system described in [5]. The system uses mixture Gaussian cross-word context dependent HMMs, each with three emitting states. Speech data is coded using 12 MFCCs, normalised log energy, and the first and second differentials of these parameters. Cepstral mean normalisation is performed on a sentence by sentence basis.

The HMMs are built in a number of stages. First, using a pronunciation dictionary and sentence orthography a phone level label string is generated by Viterbi alignment to choose the most likely pronunciation variants. These labels are used to generate single Gaussian monophone HMMs, which are then cloned for every triphone context that occurs in the training data, and the resulting single Gaussian cross-word triphone HMMs trained.

To obtain good recognition performance, mixture Gaussian densities are required, but for the majority of triphone contexts there is insufficient data to train a mixture Gaussian. Furthermore many of the cross-word triphones needed during decoding do not occur in the training data ("unseen triphones"). To solve both of these problems a tree-based state clustering procedure is used [6].

A phonetic decision tree is built for every monophone HMM state position to determine equivalence classes between sets of triphone contexts. The tree-growing procedure uses questions about the immediate phonetic context to repeatedly divide the triphones seen in training into groups. The final clusters contain triphone contexts that are acoustically similar but also have enough training observations for robust estimation of mixture Gaussians. The single Gaussian state output distributions of the members of each class are then tied to each other. By using the decision trees, the tied-state labels needed to synthesise any unseen triphones can be determined. The number of mixture components in each tied-state distribution is incremented using an iterative mixture-splitting and retraining procedure. A final optional stage in model building clones this HMM set and re-estimates separate gender-dependent mean vectors, while retaining the gender independent variances.

The system uses a time-synchronous one-pass decoder that is implemented using a dynamically built tree-structured network. This approach integrates the cross-word context-dependent acoustic models and N-gram language models directly into the search. The approach saves computation and storage by using a tree-structured lexicon since most search effort is in the first phones of each word. However it does require tree copies for different acoustic and language model contexts. Details of the decoder architecture are given in [4].

3. WORD LATTICES

Although the dynamic network decoder can operate in a single pass, for many purposes, including system development, word lattices are of great utility. A word lattice forms a compact representation of many alternative sentence hypotheses. The lattices contain a set of nodes that correspond to particular time instants and arcs connecting these nodes that represent possible word hypotheses for the time period between two nodes. Associated with each arc are both language model and acoustic model scores. Since the acoustic models include cross-word context, the lattice may contain copies of each word, and further copies can be required to encode the full language model constraint.

Once these lattices have been constructed they can be used for a number of different purposes. Language model scores can be scaled to optimise language model weights and/or word insertion penalties; a new language model can be applied using the same acoustic scores and recognition performed using an A* search through the lattice; and N-Best sentence hypotheses can be generated. Also, new acoustic models, and optionally a new language model, can be used with the lattice operating as a word-graph to constrain the search in which case the initial acoustic and language model scores and word start/end times are not required. Of course a word lattice may also form the ideal interface to pass a hypothesis set to further stages of processing.

3.1. Lattice Generation

Lattice generation only requires minor modifications to the basic search strategy since multiple copies of each word are kept during the search. Information about which words end at each node in the search at each time are recorded and retained, but only the best hypothesis is extended. The multiple hypotheses can be recovered at each word-end node at the end of the utterance. This procedure will not generate exact solutions for any but the locally best path since it implicitly assumes that the start-time of each word is independent of all words before its immediate predecessor (the "word-pair approximation"). This procedure is similar to the lattice generation described in [1] except that here it has been extended to cross-word acoustic models and lattices can be generated with arbitrary N-gram language models.

It has been found that the lattices generated in this manner can be pruned without adversely affecting lattice coverage. For each arc, the score of the best complete path through the lattice which includes that arc is found. If this score is more than a threshold from the globally best path, then the arc is deleted. This pruning strategy can be efficiently implemented using an A* algorithm.

In practice, we have found that initial lattices generated using a bigram LM are adequate. For rescoring with higher order N-gram LMs it is convenient to expand the lattice to encodes the new language model constraints along with the associated probabilities, and then apply a further stage of lattice pruning.

3.2. Lattice Error Rates

To measure the lattice quality, two lattice error rate measures are computed. The first determines whether a path corresponding to the true sentence exists in the lattice (lower bound on the sentence error rate), and the second is a lower bound on the word-error rate from rescoring the lattice, where the word-error rate is found by the usual dynamic programming string alignment procedure. Both measures are complicated by the possibility of out-of-vocabulary (OOV) words in the utterance. Lattice sentence error is only computed for sentences that don't contain OOV words, while for the lattice word error rates no account of OOV words is taken (i.e. OOV words are counted as incorrect). Therefore, if the OOV rate is subtracted from the lattice word error rate, an estimate of the lattice error rate for non-OOV words can be found.

Test Set	Density	S. Err.	OOV rate	W. Err.
Nov'92 5k	73	0.3	0.00	0.02
si_dt_05.odd	134	3.6	0.00	0.29
Nov'93 5k	135	6.5	0.29	0.73

Table 1. Lattice densities and % lattice sentence/word error rates for WSJ 5kc nvp test sets, and bigram lattices.

Test Set	Density	S. Err.	OOV rate	W. Err.
1994 H1-dev	289	16.2	0.31	1.53
Nov'94 H1	341	10.7	0.65	1.50

Table 2. Lattice quality for unlimited vocabulary test sets, 65k word vocabulary and bigram lattices.

The lattice sentence and word error rates for several test sets and systems using both 5k word and a 65k word vocabularies are shown in Tables 1 and 2. The set si_dt_05.odd contains alternate sentences from the 1993 WSJ 5k Hub development test after sentences with OOV words were removed. The other 5k test sets are from the November 1992 and 1993 5k evaluation tests. The test sets used with a 65k word recogniser are from the 1994 H1 development test and November 1994 H1 evaluation test data.

The lattices were generated with either a 5k or 20k bigram language model. The lattice density figure is the average number of lattice arcs per spoken word. As noted earlier, the lattices can contain many arcs for the same word, either with slightly different start/end times, with differing contexts, or different pronunciation variants. If the lattice arcs are merged to produce a finite-state syntax ignoring these differences the number of arcs is reduced by a factor of between 10 and 20.

Although there is some variability among the test sets, the lattice word error rates for non-OOV words is less than 0.5% for 5k test-sets and around 1% for the unlimited vocabulary data. These values imply that very similar error rates should be obtained for both rescoring the lattices and performing the full search. For acoustic rescoring the computation is reduced by approximately a factor of 20 by using the word lattices.

4. NOV'94 SYSTEM DEVELOPMENT

This section describes the the development of the system for the November 1994 ARPA CSR evaluation.

The Nov'94 ARPA evaluation H1 hub test consisted of two main parts: H1-C1 in which the acoustic training data was specified as well as a particular 20k word trigram (3-g) language model trained on 227 million words of data (CSRNAB1 text corpus) and provided by CMU; and H1-P0 in which any acoustic or language model training data could be used. In H1-C1 each sentence had to be recognised independently of any other, whereas in the H1-P0 test the speaker boundaries were known and hence it was possible to use unsupervised incremental speaker adaptation techniques. Both sets of H1 acoustic test data, (1994 development test and Nov'94 evaluation) contained about 15 sentences from 20 talkers. All results reported were scored

Vocabulary Size	System Type	GI/GD	Grammar Type	Word Penalty	Adaptation	% Word Error	
						H1-dev'94	H1 Nov'94
20k	HMM-1	GI	3-g	n	n	12.79	12.52
20k	HMM-1	GI	3-g	y	n	12.51	12.36
20k	HMM-1	GD	3-g	y	n	12.01	12.04
20k	HMM-2	GI	3-g	y	n	11.76	11.42
20k	HMM-2	GD	3-g	y	n	11.52	11.33†
65k	HMM-1	GI	3-g	y	n	9.43	9.94
65k	HMM-1	GD	3-g	y	n	9.08	9.39
65k	HMM-2	GI	3-g	y	n	9.14	9.12
65k	HMM-2	GD	3-g	y	n	8.68	8.99
65k	HMM-2	GD	4-g	y	n	8.26	8.74
65k	HMM-2	GD	4-g	y	y	7.24	8.08†

Table 3. % word error rates for different vocabulary sizes, acoustic model types, grammars etc. using the 1994 H1 development and evaluation test data. † denotes systems used for the ARPA November 1994 WSJ evaluation.

using word-mediated alignment and the preliminary reference transcriptions.

The texts that provided the prompts for the evaluation data were drawn from five different sources of North American business news from June/July 1994. The data for development test was drawn from the same sources earlier in 1994. This data was unfiltered (unlike earlier CSR material) and is hence referred to as unlimited vocabulary data.

4.1. Baseline System

Initially an HMM system was constructed using the techniques outlined in Sec. 2. The system described in [5] used the Dragon Wall Street Journal Pronunciation Lexicon version 2.0 with some local modifications and corrections. However the new baseline system was built using the 1993 LIMSI WSJ Lexicon and phone set. Preliminary experiments showed that on average this change gave about a 4% word error rate reduction.

The acoustic training data used for both the baseline system and all subsequent HMM sets consisted of 36493 sentences from the SI-284 WSJ0+1 data set. These data were used to build a gender independent (GI) triphone HMM set with 6399 speech states, with each state having a 12 component Gaussian mixture output distribution. This HMM set is referred to as HMM-1, and the performance of this system with the standard CMU 20k trigram grammar is given in the first line of Table 3. The word lattices used for all results in Table 3 were generated using the GI HMM-1 set with the appropriate bigram LM.

4.2. Word-Insertion Penalty

The baseline system used a grammar scale factor but without a word-insertion penalty. With word lattices it is very quick to compute the optimal values for both the grammar scale and word insertion penalty, although in fact the system has been found to be fairly insensitive to the exact values used. The first two lines of Table 3 show that the use of a word insertion penalty decreased the word error rate by about 2%.

4.3. Improved Acoustic Models

The tree-clustering used in constructing HMM-1 the set considers only the immediate neighbouring phonetic context (triphone context). To investigate the use of wider phonetic contexts a model set was constructed in which the tree clustering could ask questions about the preceding and following two phones (quinphone context) and also the position of word boundaries. To facilitate this a new implementation of the tree clustering software was required that directly used state-level Viterbi alignments of the training data to compute the necessary statistics rather than first building a set of single Gaussian context dependent HMMs and saving the state occupation counts. The model set with quinphone context (denoted HMM-2) had 9354 speech states, and each state characterised by a 14 component mixture Gaussian. Hence HMM-2 contained about 1.7 times as many parameters as HMM-1. Taking into account all the contrasts in Table 3, it can be seen that on average the decrease in error rate due to HMM-2 is about 5%.

Gender dependent (GD) versions of both HMM-1 and HMM-2 were built. In each case the gender independent system was cloned and the means and mixture weights of each set retrained using a single iteration of Baum-Welch re-estimation with only the training data from one gender. The variances were left fixed at their gender independent values and tied to save memory. Comparing results for the corresponding GD and GI systems in Table 3 it can be seen that gender dependent models improve performance by about 3%.

System	Nov'92	si_dt_20.odd	Nov'93
From [5]	9.46	13.71	12.74
HMM-2 GD	8.19	12.34	11.61

Table 4. % word error rates comparing the GD SI-284 system from [5] and GD HMM-2 with word insertion penalty.

A comparison of the HMM-2 system (quinphone context, LIMSI dictionary & word insertion penalty) with the SI-284 GD system described in [5] is shown in Table 4. Nov'92 refers to the 1992 20k nvp evaluation test set, si_dt_20.odd refers to alternate sentences in the 1993 H1 development test data and Nov'93 to the 1993 H1 evaluation test data. Note that for these tests the 1993 20k word list and Lincoln Labs 20k trigram grammar were used. The HMM-2 system reduces the word error rate by about 11%, and this system gave the lowest error rate in the H1-C1 test of 1994 ARPA evaluation.

4.4. 65k Word Vocabulary

If the CMU 20k word-list is used, 2.68% of words are OOV in the 1994 H1-dev and 2.38% in the Nov'94 H1 data. Since on average 1.6 errors occur for each OOV word in the test data, OOV words have a significant impact on the overall recognition rate. Therefore to reduce the OOV rate, a recogniser with a 65464 (65k) word vocabulary was developed. The word-list was chosen by taking the most frequent words in the CSRNAB1 corpus and filtering to remove mis-spelled words etc. and then adding the the most frequent extra words from the 1.4 million word 1994 LM development-test text corpus. This procedure resulted in an OOV rate of 0.31% on the acoustic development test data and 0.65% on the evaluation data.

The additional pronunciation entries needed by the recogniser for a 65k vocabulary were generated using a text-to-speech system and then those entries were mapped to the LIMSI phone set.

A 65k backoff trigram [3] language model was estimated from the CSRNAB1 training texts and the development test data. The LM used for the development test data excluded the articles from which the prompts for the acoustic data was taken, but the LM used for the Nov'94 evaluation test included all articles. To model the differences between the spoken data and the canonical form of the text data, particularly for numbers, some of the text data were modified to reflect the forms seen in the acoustic training data.

Comparing the entries for 20k and 65k vocabularies in Table 3, the error rate reduction was roughly 24% with the 65k trigram on the development test and 20% on the evaluation test—the difference being largely due to the difference in OOV rate reduction. It was found that about a 0.25% (absolute) reduction in error rate was due to modelling the effect of spoken numbers, and that 0.4% (absolute) was gained from adding the development texts on the development data but only 0.1% on the evaluation data.

4.5. 4-Gram Language Model

The use of a backoff 4-gram (4-g) LM was investigated to see if the extra span would be beneficial with a large amount of training data. The test-set perplexity of the 65k 3-g and 4-g LMs are given in Table 5. The 65k 4-g LM contained 6.3 million bigrams, 9.9 million trigrams and 10.1 million 4-grams. The 4-gram yields a 9% perplexity reduction and Table 3 shows that it resulted in an average decrease in word error rate of 4%.

Test Set	% OOV	3-g perplexity	4-g perplexity
1994 H1-dev	0.31	145.2	132.7
Nov'94 H1	0.65	145.3	131.8

Table 5. Test set perplexities for 65k 3-g and 4-g LMs

4.6. Speaker Adaptation

Although speaker independent systems can have good overall performance, some speakers are poorly modelled by such a system. If a sequence of utterances are from the same speaker then the system may be able to be adapted to better model the current speaker. This process should ideally be *unsupervised* (i.e. the true transcription for each adaptation utterance is not required), and *incremental* so that performance will gradually improve as more adaptation data is available. The GD HMM-2 HMM set contains around 15 million parameters, and a key issue is adapting the parameters of such a system using a small amount of adaptation data. The approach adopted here is an extension of the maximum likelihood linear regression (MLLR) technique [2].

MLLR estimates the parameters of a set of full (40×39) transformation matrices which are used to transform the Gaussian mean vectors. The matrices are estimated so as to maximise the likelihood of the transformed models generating the adaptation data. The technique is implemented using the forward-backward algorithm and has close links with standard Baum-Welch training.

When only a small amount of data is available all Gaussian means are transformed by the same matrix (single class), and as more data becomes available more specific matrices are computed using only the data that is aligned with that class. The number of classes used at any stage depends on the available adaptation data. All the Gaussian means are clustered into a set of 750 base classes, these are then arranged into a heirarchy and the most specific class is generated that has enough observations to robustly estimate the MLLR matrix parameters.

The adaptation was implemented in an acoustic rescoring pass using the 4-g word lattices. The gender of the current speaker was found from the first two utterances, and then only the model set for the recognised gender adapted. The model means were updated after every two sentences. It can be seen from the final two lines of Table 3 that MLLR speaker adaptation reduced the error rate between 8% and 12%, and moreover had the greatest effect on the speakers that were most poorly modelled by the original system.

The 65k GD HMM-2 4-g system with unsupervised incremental speaker adaptation gave the lowest error rate in the H1-P0 test of the 1994 ARPA evaluation.

5. CONCLUSION

The use of word lattices in the HTK large vocabulary recognition system has enabled a number of new techniques to be investigated including wider context phonetic models, 4-gram language models and unsupervised speaker adaptation. The resulting system has been shown to give state-of-the-art performance.

ACKNOWLEDGMENTS

This work is in part supported by an EPSRC/MOD research grant (GR/J10204) and EPSRC grant reference GR/K25380. C.J. Leggetter and J.J. Odell are supported by EPSRC research studentships.

REFERENCES

[1] Aubert X., Dugast C., Ney H. & Steinbiss V. (1994). Large Vocabulary Continuous Speech Recognition Of Wall Street Journal Data. *Proc. ICASSP'94*, Vol. 2, pp. 129-132, Adelaide.

[2] Leggetter C.J. & Woodland P.C. (1994). Speaker Adaptation of Continuous Density HMMs Using Linear Regression. *Proc. ICSLP'94*, Vol. 2, pp. 451-454, Yokohama.

[3] Katz S.M. (1987). Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recogniser. *IEEE Trans. ASSP*, Vol. 35, No. 3, pp. 400-401.

[4] Odell J.J., Valtchev V., Woodland P.C. & Young S.J. (1994). A One Pass Decoder Design For Large Vocabulary Recognition. *Proc. ARPA Human Language Technology Workshop, March 1994*. Morgan Kaufmann.

[5] Woodland P.C., Odell J.J., Valtchev V. & Young S.J. (1994). Large Vocabulary Continuous Speech Recognition Using HTK. *Proc. ICASSP'94*, Vol. 2, pp. 125-128, Adelaide.

[6] Young S.J., Odell J.J. & Woodland P.C. (1994). Tree-Based State Tying for High Accuracy Acoustic Modelling. *Proc. ARPA Human Language Technology Workshop, March 1994*. Morgan Kaufmann.