# THE HTK LARGE VOCABULARY RECOGNITION SYSTEM FOR THE 1995 ARPA H3 TASK

*P.C. Woodland, M.J.F. Gales, D. Pye & V. Valtchev*

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

## ABSTRACT

The HTK large vocabulary speech recognition system has previously shown very good performance for clean speech. This paper describes developments of the system aimed at recognition of speech from the ARPA H3 task which contains data of a relatively low signal-to-noise ratio from unknown microphones. It is shown that a two-phase approach can be effective. The first phase is to derive an initial set of models that are more appropriate for the current conditions than using models trained on clean speech. This is done using either single-pass retraining with multiple microphone data or parallel model combination which combines HMMs trained on clean data with estimates of convolutional and additive noise. The second stage provides more detailed environmental and speaker adapatation using maximum likelihood linear regression which estimates a set of linear transformations of the model parameters to the current conditions. Experiments are reported on both the 1994 ARPA CSR S5 (alternate microphones) and S10 (additive noise) spoke tasks as well as the 1995 ARPA CSR H3 task. The HTK system yielded the lowest error rates in both the H3-P0 and H3-C0 tests.

## 1. INTRODUCTION

This paper describes the development of the HTK LVCSR system, which gives state-of-the-art performance under clean speech conditions, [11], to non-ideal acoustic environments, with a focus on recognition of data from the 1995 ARPA H3 task. This data has a relatively low signal-to-noise ratio and is also from unknown microphones.

A two-phase approach is adopted to producing a system appropriate for the current environmental conditions. Firstly, an initial set of models are derived that should be more appropriate than the clean model set for the environmental conditions under test. This first stage uses either single-pass retraining (SPR) [2] with data from multiple environments or uses parallel model combination (PMC) [1, 2] which combines estimates of convolutional and additive noise to compensate an HMM set trained on clean speech.

The second stage requires more data from the new environment but can produce more detailed environmental compensation along with speaker adaptation. This phase uses maximum likelihood linear regression (MLLR) [5, 6] which, in its original form, estimates a set of linear transformations for the Gaussian mean parameters. Recently [3] we have extended the MLLR approach so that the Gaussian variance parameters can also be compensated. While MLLR produces good results with only a modest amount of data from the new speaker/environment, if more data is available it is able to provide more precise adaptation.

This paper gives an overview of the HTK LVCSR system and briefly describes the approaches to deriving the initial models for a new environment. A description of MLLR including the extensions for variance compensation is then given. The approach for environment compensation is evaluated first using the 1994 ARPA CSR S5 (alternate microphones) and S10 (additive noise) spoke tasks. Finally the HTK system for 1995 ARPA Hub 3 evaluation is described in detail.

## 2. CLEAN SPEECH SYSTEM

This section gives an overview of the clean-speech HTK LVCSR system. The system uses state-clustered, cross-word mixture Gaussian context-dependent acoustic models and a back-off N-gram language model. More details of the system can be found in [11].

In the standard system, each speech frame is represented by a 39 dimensional feature vector that consists of 12 mel frequency cepstral coefficients, normalised log energy along with the first and second differentials of these values. Cepstral mean normalisation (CMN) is applied. For use with PMC, the front end is slightly modified: the zeroth cepstral coefficient replaces log energy; no CMN is performed and the regression-smoothed differentials are replaced by simple differences. We have also investigated the use of a PLP-based [4] cepstral parameterisation (see Secs. 5 and 6).

The HMMs are built in a number of stages. First, the LIMSI 1993 WSJ pronunciation dictionary is used to generate phone level labels for the training data. Then in turn single Gaussian monophone HMMs, single Gaussian cross-word triphone models and single Gaussian state-clustered triphones are trained. The clustering is decision-tree based [13] to allow for the synthesis of triphone models that don't occur in training. After clustering mixture Gaussians are estimated by iterative "mixture-splitting" and forward-backward retraining.

The acoustic training for the clean-speech system consisted of 36,493 sentences from the SI-284 WSJ0+1 data sets. These data were used to build a gender independent triphone HMM set with 6,399 speech states, with each state having a 12 component Gaussian mixture output distribution. This system (the HMM-1 system of [11]) was used as the basis for the S5 and S10 experiments.

The full HTK LVCSR system also uses more complex acoustic models which take account of the preceding and following

two phones (quinphone context) and also the position of word boundaries. The gender independent version of this HMM set (the HMM-2 system of [11]) had 9,354 speech states with each state characterised by a 14 component mixture Gaussian. Gender dependent versions of this system are trained by using the data from just the relevant training speakers and updating the means and mixture weights.

The HTK LVCSR system uses a time-synchronous decoder employing a dynamically built tree structured network decoder [8]. This decoder can either operate in a single pass or it can be used to produce word lattices which compactly store multiple sentence hypotheses. The lattices contain both language model and acoustic information and can be used for rescoring with new acoustic models, or for the application of new language models.

# 3. INITIAL MODEL DERIVATION

This section describes two methods for providing initial models that better match a new environment than clean speech models. Both methods require very little data from the new environment.

## 3.1. Single-Pass Retraining

Single pass retraining operates by use of a stereo database in which there are paired speech samples, one clean and the other for an approximation to the new environment ("secondary channel"). Given a mixture Gaussian HMM system trained on clean speech, and assuming that the frame/state (Gaussian mixture component) alignment is identical for the clean and the secondary channel data, SPR first finds the *a posteriori* probability of mixture component occupation using the clean speech models and the clean speech vectors. Using this clean speech alignment, the Gaussian parameters are updated using the corresponding observations from the secondary channel.

SPR is here used to train an initial system from stereo data that contains multiple second channel microphones. This gives a system that is better matched to the new environment than using a clean model set and serves as a suitable initialisation for further MLLR-based adaptation.

## 3.2. Parallel Model Combination

PMC attempts to estimate the parameters of a matched HMM system given the clean speech models, a model of additive interfering noise and the frequency response of the channel difference between clean speech training conditions and the test environment. It is assumed that speech and noise are independent and additive in time and (linear) frequency domains and that a Gaussian or mixture Gaussian model is sufficient to describe the noise process in the log spectral or cepstral domains. Although HMM modelling is performed in the cepstral domain, compensation is performed in the linear spectral and log spectral domains by using the appropriate transformations.

There are a number of different PMC approximations and implementations that have been investigated [2]. However one of the simplest and most efficient is Log-Add PMC which updates just the mean HMM parameters. This method essentially assumes that that the speech and noise models have zero variance. If a compensated Gaussian mean component in the log spectral domain is denoted as $\hat{\mu}_i$ then

$$\hat{\mu}_i = \log(\exp((H_i + \mu_i) + \exp(\tilde{\mu}_i))$$

where $H_i$ is the channel difference between training and the new environment, $\mu_i$ is the clean speech mean and $\tilde{\mu}_i$ the noise mean in the log spectral domain. The means of the 1st and 2nd differentials can be compensated in a similar way.

# 4. MAXIMUM LIKELIHOOD LINEAR REGRESSION

MLLR was originally developed for speaker adaptation [5, 6] but can equally be applied to situations of environmental mismatch. MLLR estimates a set of transformation matrices for the HMM Gaussian parameters which maximises the likelihood of the adaptation data. The set of transformations is relatively small compared to the total number of Gaussians in the system and so a number of Gaussians share the same transformation matrices. This means that the transformation parameters can be robustly estimated from only a limited amount of data which allows all the Gaussians in the HMM set to be updated. For a small amount of data (or very robust transformation estimation) only a single global transformation is used. As more data becomes available more specific transformations can be estimated.

Originally transformations were estimated only for the mean parameters but we have recently extended the approach so that the Gaussian variances can also be updated [3]. In the systems described here, MLLR is used to fine-tune a system to the environment/speaker after either the application of PMC or use of SPR with secondary channel data. This section gives a brief overview of the basic MLLR theory for both the mean parameters and the variances.

The means and variances are adapated in two separate stages. Initially new means are found and then given these new means the variances are also updated. The HMMs are modified such that

$$\mathcal{L}(\mathbf{O}_T|\check{\mathcal{M}}) \geq \mathcal{L}(\mathbf{O}_T|\hat{\mathcal{M}}) \geq \mathcal{L}(\mathbf{O}_T|\mathcal{M})$$

where for the models $\hat{\mathcal{M}}$ have just the means parameters updated to $\hat{\mu}_1, \ldots, \hat{\mu}_M$; the models $\check{\mathcal{M}}$ have both the means and the variances $\hat{\mathbf{\Sigma}}_1, \ldots, \hat{\mathbf{\Sigma}}_M$ updated and $\mathbf{O}_T$ is the adaptation data.

$$\mathbf{O}_T = \{\mathbf{o}(1), \mathbf{o}(2), \ldots, \mathbf{o}(T)\}$$

## 4.1. MLLR Adaptation of the Means

The aim of MLLR is to obtain a set of transformation matrices that maximises the likelihood of the adaptation data. The transformation matrix is used to give a new estimate of the mean, where

$$\hat{\mu}_m = \hat{\mathbf{W}}_m \xi_m$$

and $\hat{\mathbf{W}}_m$ is the $n \times (n+1)$ transformation matrix (for $n$ dimensional data) and $\xi_m$ is the extended mean vector

$$\xi_m = \begin{bmatrix} 1 & \mu_1 & \ldots & \mu_n \end{bmatrix}^T$$

In order to ensure robust estimation of the transformation parameters the transformation matrices are tied across a number of Gaussians according to a regression class tree [6]. This tree contains all the Gaussians in the system and statistics are gathered at the leaves (which may each contain a number of Gaussians). The most specific transform that can be robustly estimated is then generated for all the Gaussians in the system.

A particular transformation $\hat{\mathbf{W}}_m$ is to be tied across $R$ components $\{m_1, \ldots, m_R\}$. For the Gaussian output probability density functions considered, $\hat{\mathbf{W}}_m$ may be found by solving

$$\sum_{\tau=1}^{T}\sum_{r=1}^{R} L_{m_r}(\tau)\boldsymbol{\Sigma}_{m_r}^{-1}\mathbf{o}(\tau)\xi_{m_r}^{T} = \sum_{\tau=1}^{T}\sum_{r=1}^{R} L_{m_r}(\tau)\boldsymbol{\Sigma}_{m_r}^{-1}\hat{\mathbf{W}}_m\xi_{m_r}\xi_{m_r}^{T}$$

where

$$L_m(\tau) = p(q_m(\tau)|\mathcal{M}, \mathbf{O}_T)$$

and $q_m(\tau)$ indicates Gaussian component $m$ at time $\tau$. For the full covariance matrix case the solution is computationally very expensive [3]. The solution for the diagonal covariance case is computationally tractable and described in [5]. Each transformation can be a full matrix or constrained to be block diagonal or diagonal.

## 4.2. MLLR Variance Adaptation

The Gaussian variance vectors, or in general covariance matrices, are updated using the following transformation

$$\hat{\boldsymbol{\Sigma}}_m = \mathbf{B}_m^T\hat{\mathbf{H}}_m\mathbf{B}_m$$

where $\hat{\mathbf{H}}_m$ is the linear transformation to be estimated and $\mathbf{B}_m$ is the inverse of the Choleski factor of $\boldsymbol{\Sigma}_m^{-1}$, so

$$\boldsymbol{\Sigma}_m^{-1} = \mathbf{C}_m\mathbf{C}_m^T$$

and

$$\mathbf{B}_m = \mathbf{C}_m^{-1}$$

In a similar fashion to the means the variances transformation is shared over a number of components, $\{m_1, \ldots, m_R\}$. It is simple to show that the maximum likelihood estimate is

$$\hat{\mathbf{H}}_m = \frac{\sum_{r=1}^{R}\mathbf{C}_{m_r}^T\left[\sum_{\tau=1}^{T}L_{m_r}(\tau)(\mathbf{o}(\tau)-\hat{\mu}_{m_r})(\mathbf{o}(\tau)-\hat{\mu}_{m_r})^T\right]\mathbf{C}_{m_r}}{\sum_{r=1}^{R}\sum_{\tau=1}^{T}L_{m_r}(\tau)}$$

where $\hat{\mu}_{m_r}$ is the mean previously calculated. It can be seen that the variance transformation matrix will be full, yielding full covariance matrices for each component. A diagonal transformation for the variances may be obtained by simply zeroing the off-diagonal terms and this is still guaranteed to increase the likelihood. A diagonal variance transformation is used in all the experiments reported in this paper.

The mean transformation matrix is a function of the component variance. Thus by altering the variance, the maximum likelihood estimate of the mean transformation will also be altered. While an iterative scheme could be used it has been found that a single update of the means and variances is, in practice, sufficient.

# 5. S5/S10 EXPERIMENTS

In this section the performance of the above techniques on the 1994 ARPA CSR Spoke 5 (S5) and Spoke 10 (S10) evaluation data sets is explored. S5 and S10 are 5k word tasks and use a standard 5k trigram language model. All results have been generated using the official NIST scoring software. In all cases, for ease of experimentation, the model sets used are based on the HMM-1 set and the more sophisticated HMM-2 model sets were not used. All results quoted are the result of full decoding passes and do not use word lattices generated by different systems. Further details of these experiments can be found in [12].

In all systems that use initial PMC compensation, the channel mismatch, $H_i$, was estimated in the manner described in [1] using a 30-component Gaussian mixture model and the first sentence from each speaker.

## 5.1. S10 Experiments

S10 concerns additive noise: the test data consists of clean data with car noise added at different overall SNRs. The experiments here use the S10 level 3 evaluation data which had an A-weighted SNR of 10dB, which was the lowest SNR available. The data consists of 113 sentences from 10 speakers. PMC used a noise model built using the background noise sample provided with the dataset.

The performance using the uncompensated clean models on this data is very poor giving a 54.3% word error rate, while the the error rate on the corresponding clean data with the standard MFCC parameterisation is 5.8% and 6.7% with that used with PMC. The use of Log-Add PMC reduces the error rate to 10.7%.

| Updating Means | Updating Variances | Word Error (%) |
|---|---|---|
| × | × | 10.7 |
| √ | × | 9.3 |
| √ | √ | 8.9 |

Table 1: Unsupervised incremental MLLR on PMC Log-Add S10 system.

Table 1 shows that the use of MLLR further reduces the error rate of the PMC-compensated system. Note that MLLR is being applied in incremental adaptation mode and therefore only the final sentence for each talker will gain full benefit. The incremental unsupervised adaptation reduces the error rate by about 17% with variance adaptation contributing 4%.

These systems would have been valid S10 evaluation systems. The best result (8.9 %) represents a 27% lower error rate than the best value reported in the 1994 evaluation for S10 level 3 (12.2% [9]).

## 5.2. S5 Experiments

The S5 data consisted of 200 sentences from 20 speakers. For each speaker one of 10 alternate microphones was used.

The set included tie-clip microphones, stand-mounted microphones and a hand-held microphone. The A-weighted SNR was typically 20dB.

In this case, PMC-compensated models and models trained on secondary channel data using SPR were compared as the initial models for incremental MLLR. The secondary channel data is from the SI-284 training set and it was recorded using a selection of 13 different microphones and low noise conditions. None of the microphones used for the SI-284 secondary channel data are of the same type as used in the test data. Preliminary investigations had shown that a perceptual linear prediction (PLP) [4] speech parameterisation was more robust to mismatched environments than standard MFCCs, so a PLP-based secondary channel version of the HMM-1 set was trained by SPR.

For the S5 PMC-based experiments, both the channel distortion and the background noise was estimated using the first sentence from each speaker.

| Model Set | Baseline | Incremental MLLR | |
|---|---|---|---|
| | | Means | Means+Vars |
| Clean | 17.4 | 12.1 | — |
| PMC Log-Add | 10.3 | 8.6 | 8.0 |
| PLP 2nd channel | 9.0 | 7.4 | 7.1 |

Table 2: % Word error rates for S5 data.

Table 2 shows that if clean models are used the channel distortion causes a large increase in error. It should be noted that even though the standard system includes CMN, in the presence of background noise it is not particularly effective. Both PMC and particularly the PLP 2nd channel system reduce the error rate significantly over the clean model result. Incremental MLLR adaptation again provides improvements with variance compensation further decreasing the error rate by about 5%. The use of incremental adaptation means that the results would have been valid as S5 evaluation systems. The 7.1% word error rate for the secondary channel system with mean and variance MLLR compares favourably to the best 1994 S5 evaluation result (9.7% [9]).

# 6. NOV'95 H3 EVALUATION SYSTEM

This section describes the H3 test and the HTK system used for the 1995 H3 evaluation. The ideas developed in the previous sections were built upon and multiple iterations of adaptation performed in a number of separate passes through the test data. The detailed results of each of these passes is given as well as information on the language models used.

## 6.1. Test Data

The Nov'95 ARPA H3 task was to recognise speech data read from US newspaper articles published in August 1995. The data was not filtered (unlimited vocabulary test). The speech was collected in a noisy environment with simultaneous recording from a number of far-field microphones as well as a close-talking microphone. For each speaker one far-field microphone was chosen as the test material for H3-P0, and the same speech captured by the close-talking microphone used for the H3-C0 test. Each of 20 speakers read 15 sentences from one news article. The test was defined so that data for each speaker (or session) could be processed as a block ("transcription mode"). This permits multiple unsupervised adaptation passes through the data. The A-weighted SNR of the H3-P0 data from each speaker varied from about 7dB to 23dB.

## 6.2. HTK H3 System

The HTK system developed for the tests had two paths: one for high SNR signals typical of the H3-C0 data and one for low SNR data typical of the H3-P0 data. First the data for a session was classified as either high or low SNR and then processed accordingly. Both paths included similar processing: the main difference being that the HMMs used for high SNR were trained using the Sennheiser SI-284 training data and the low SNR data used models trained using the secondary channel data. Gender independent versions of both HMM-1 and HMM-2 [11] systems were trained for both paths using the PLP representation by SPR from the corresponding clean MFCC based systems. Furthermore gender dependent HMM-2 high SNR models were also trained.

The language models were trained on a total of 406 million words of text from the 1995 reprocessed CSRNAB1 text training corpus, the 1994 development text corpus, and the H3 and H4 text data sets. All texts predated August 1 1995. For the H3 and H4 texts 889 additional abbreviations were expanded in the text training data. A word list with 65,478 entries was derived from the most frequent words used in a subset of the data and back-off bigram, trigram and 4-gram language models built [7]. The OOV rate of the test data (accounting for the official mappings used in scoring) was 0.56%. Pronunciation information came from the LIMSI 1993 WSJ Lexicon augmented with pronunciations generated by a text-to-speech system, along with some hand-generated corrections.

Decoding operated on a session by session basis in a number of stages. All stages used the dynamic network decoder [8] which allows single-pass decoding, lattice generation and lattice constrained decoding. All adaptation stages compensated both means and variances by MLLR and used block diagonal MLLR matrices for the means. This was found to be more robust than the use of full matrices when multiple unsupervised adaptation passes are used.

First, two preliminary passes were performed on the data using the HMM-1 models with tight pruning to give a rough initial transcription. The first of these used the original models and the second uses global MLLR adaptation (i.e. a single transformation for all Gaussians) and the trigram language model. Using the transcriptions from the second preliminary pass, global MLLR adaptation was again performed. These models were used to generate word lattices using a bigram language model. The use of these preliminary passes had been found to be vital to generate high quality lattices suitable for subsequent recognition passes.

The bigram lattices were expanded to trigram and using the HMM-1 models with more specific MLLR adaptation, the final HMM-1 output was derived. This was then used to adapt the HMM-2 models using 4-gram lattices.

For the high SNR path, the gender of HMM-2 models for subsequent passes was found using the likelihoods from forced alignments of the final HMM-1 output with the male and female model sets—gender independent models were used if there was inconsistency within a session.

Finally the 4-gram lattices were iteratively rescored using the HMM-2 models. The final HMM-1 transcriptions and global adaptation (with a separate transform for silence) were initially used and then on each subsequent iteration a larger number of regression classes were created. There were 5 such HMM-2 passes for the low-SNR data and 3 passes for the high SNR data. The final pass gave the system output.

## 6.3. Evaluation System Results

Table 3 shows the scored output of the system at various stages of processing. It can be seen that there is a substantial decrease in word error rate between the first two preliminary passes (Prelim. 1 and Prelim. 2) which leads to a much improved lattice word error rate in the lattice generation stage. The final HMM-1 output uses a number of transformation matrices (the previous stages use global adaptation). If this stage had been the final output of the system both the H3-P0 and H3-C0 systems would have given the lowest error rates in the Nov'95 H3 evaluation. The use of the HMM-2 set of models along with the 4-gram language model decreases the error rate by about a further 15%. The last line of Table 3 gives the actual HTK results in the Nov'95 H3 evaluation which were the lowest error rates in both the H3-P0 and H3-C0 tests. All the results use the adjudicated transcriptions and map files.

| Processing Stage | LM Type | H3-P0 Data | H3-C0 Data |
|---|---|---|---|
| Prelim. 1 | tg | 33.27 | 12.59 |
| Prelim. 2 | tg | 21.06 | 9.60 |
| Lattice Gen. | bg | 22.12 | 10.88 |
| Lattice Gen. | tg | 17.20 | 7.88 |
| Final HMM-1 | tg | 16.17 | 7.61 |
| Global HMM-2 | fg | 14.49 | 6.81 |
| HMM-2 thresh. a | fg | 14.24 | — |
| HMM-2 thresh. b | fg | 13.81 | — |
| HMM-2 thresh. c | fg | 13.71 | 6.68 |
| Final HMM-2 | fg | 13.50† | 6.63† |

Table 3: % Word error rates on Nov'95 H3 data at various stages of processing. † denotes the systems actually used for the Nov'95 H3 evaluation.

There are a number of stages of processing with the HMM-2 model sets and in each step the number of transformation matrices is increased. The decrease in word error using multiple transformations with the HMM-2 models on the H3-P0 data is 7%—this becomes just a 3% reduction (to 14.11%) if the intermediate stages of adaptation are not performed.

## 6.4. Effect of Adaptation Type

Table 3 shows the result of using the HMM-2 models with the 4-gram evaluation lattices with either no adaptation or mean-only MLLR. Although the lattices were derived using mean and variance MLLR, we expect the figures to be an accurate estimate of the error rate since the lattices are large. Also it should be noted that the grammar scale and word-insertion penalties were not tuned for these contrasts. The use of mean and variance adaptation gives a large decrease in error rate: 39% on H3-P0 and 22% on H3-C0 data, while mean adaptation alone produces reductions of 31% and 17% respectively. These percentage decreases in word error rates due to MLLR for the H3-P0 data are nearly double those given for the S5 data because of the increased mismatch between the secondary channel training data and the H3 test data and also the use of multiple iterations of transcription mode adaptation. Variance adaptation is particularly important for noisy data since noise reduces the speech variance. Mean and variance MLLR provided a fairly consistent improvement (relative to mean MLLR) across speakers: for both H3-P0 and H3-C0 only 2 speakers gave more errors with the addition of variance adaptation.

| Adaptation | H3-P0 Data | H3-C0 Data |
|---|---|---|
| None | 22.12 | 8.54 |
| Means | 15.22 | 7.11 |
| Means+Vars | 13.50 | 6.63 |

Table 4: % Word error rates on Nov'95 H3 data with different types of adaptation.

## 6.5. PMC Initial Models

The actual evaluation system used initial models trained by SPR on secondary channel data and a PLP data parameterisation. Although it wasn't feasible to re-run the entire evaluation system using PMC initial models the two preliminary passes were re-run using this data. The noise estimates were obtained from the noise samples provided. The results are given in Table 5. It can be seen that the PLP secondary channel system has about 11% fewer errors for the initial pass and 8% fewer errors for the second preliminary pass. It would be expected from these results that if the complete system was run that the difference between the PLP based system and a PMC-based one (which has the advantage of not needing secondary channel data) would be rather less than 10%.

| Stage | PMC-based | 2nd channel |
|---|---|---|
| Prelim. 1 | 37.54 | 33.27 |
| Prelim. 2 | 22.98 | 21.06 |

Table 5: % Word error for the two H3-P0 preliminary passes with either secondary channel or PMC initial models

## 6.6. Effect of LM Training Data

After the evaluation a set of perplexity measurements were made on the financial data from August 1995, a subset of which was used as the evaluation data prompts. The perplexities of LMs trained using all the available training data and various subsets of the data were compared. These measurements are shown in Table 6. All these language models used the same word list which had an OOV rate of 0.7%, and used the same cut-off values. It can be seen that the addition

| LM Training Data | bigram | trigram | 4-gram |
|---|---|---|---|
| NAB1+dev94 | 222 | 141 | 129 |
| NAB1+dev94+H4 | 221 | 136 | 122 |
| NAB1+dev94+H3 | 203 | 121 | 107 |
| NAB1+dev94+H3+H4 | 206 | 120 | 105 |

Table 6: Test set perplexities for language models trained on different data subsets.

of the H3 data to the older data (NAB1 and dev94) produces worthwhile reductions in the perplexity but the addition of the H4 data was much less useful. It should be noted that, as expected, as more data is added the perplexity reduction due to the 4-gram language model increases.

## 6.7. Cache Language Models

The final system lattices have been rescored using an adaptive language model. For each utterance, the previous best hypothesis of each of the remaining fourteen sentences in that session were used to construct rare-word unigram and bigram caches of the type described by [10]. The set of words in the unigram cache was supplemented by words sharing the same stem. The recognition output was produced using an A* search of the lattices from the final acoustic pass with LM probabilities from the caches (updated with the sentence hypothesis so far) dynamically interpolated with the standard N-Gram LM scores. It can be seen from Table 7 that the use

| Cache | H3-P0 Data | H3-C0 Data |
|---|---|---|
| No | 13.50 | 6.63 |
| Yes | 13.27 | 6.42 |

Table 7: % Word error rates on Nov'95 H3 data with and without using a final cache language model pass.

of a cache has improved the word error rates by about 0.2% absolute (an average of 2% relative improvement). The improvement is limited by the quantity of LM adaptation data and the fact that the final lattices are rather small. However the results still represent the lowest error rates that we have achieved on the H3 data.

# 7. CONCLUSION

The development of the HTK system for the 1995 ARPA H3 evaluation has been described. A combination of techniques have been shown to be effective and the resulting system gives state-of-the-art performance on data with both additive noise and channel effects.

# 8. ACKNOWLEDGEMENTS

This work is in part supported by an EPSRC grant reference GR/K25380. Mark Gales is supported by a Research Fellowship from Emmanuel College, Cambridge. Additional computing resources were provided by the ARPA CAIP computing facility.

# References

1. Gales M.J.F. & Young S.J. (1995). Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination. *Computer Speech & Language*, Vol. 9, pp. 289-308.
2. Gales M.J.F. (1996) Model-Based Techniques for Noise Robust Speech Recognition. Ph.D. Thesis, Cambridge University.
3. Gales M.J.F. & Woodland P.C. (1996). Variance Compensation Within the MLLR Framework. Technical Report CUED/F-INFENG/TR.242. Cambridge University Engineering Department.
4. Hermansky H. (1990). Perceptual Linear Prediction (PLP) Analysis for Speech. *J. Acoust. Soc. Amer.*, Vol. 87, pp. 1738-1752.
5. Leggetter C.J. & Woodland P.C. (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech & Language*, Vol. 9, pp. 171-185.
6. Leggetter C.J. & Woodland P.C. (1995). Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression. *Proc. ARPA Spoken Language Technology Workshop*, pp. 104-109. Morgan Kaufmann.
7. 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.
8. 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*, pp. 405-410. Morgan Kaufmann.
9. Pallett D.S. et al. (1995). 1994 Benchmark Tests for the ARPA Spoken Language Program. *Proc. ARPA Spoken Language Technology Workshop*, pp. 5-36. Morgan Kaufmann.
10. Rosenfeld R. (1994) Adaptive Statistical Language Modeling: A Maximum Entropy Approach. Ph.D. Thesis, Carnegie-Mellon University.
11. Woodland P.C., Leggetter C.J., Odell J.J., Valtchev V. & Young S.J. (1995). The 1994 HTK Large Vocabulary Speech Recognition System. *Proc. ICASSP'95*, Vol. 1, pp. 73-76, Detroit.
12. Woodland P.C., Gales M.J.F. & Pye D. (1996). Improving Environmental Robustness in Large Vocabulary Speech Recognition. To appear *Proc. ICASSP'96*, Atlanta.
13. 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*, pp. 307-312, Morgan Kaufmann.