ITERATIVE UNSUPERVISED ADAPTATION USING MAXIMUM LIKELIHOOD LINEAR REGRESSION

P.C. Woodland *D. Pye* *M.J.F. Gales*

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

ABSTRACT

Maximum likelihood linear regression (MLLR) is a parameter transformation technique for both speaker and environment adaptation. In this paper the iterative use of MLLR is investigated in the context of large vocabulary speaker independent transcription of both noise free and noisy data. It is shown that iterative application of MLLR can be beneficial especially in situations of severe mismatch. When word lattices are used it is important that the lattices contain the correct transcription and it is shown that global MLLR based on rough initial transcriptions of the data can be very useful in generating high quality lattices. MLLR can also be used in an iterative fashion to refine the transcriptions of the test data and adapt models based on the current transcriptions. These techniques were used by the HTK large vocabulary system for the November 1995 ARPA H3 evaluation. It is shown that iterative application MLLR prior to lattice generation and for iterative refinement proved to be very effective.

1. INTRODUCTION

One of the most important issues for large vocabulary speaker independent continuous speech recognition systems is recognition of data that is not well represented in the system training set. This includes recognition of outlier speakers and particularly speech recorded in mismatched conditions through additive noise or a different microphone.

The general approach that we have adopted for recognition of mismatched speech is to first derive "initial" models that are more appropriate to the conditions under test than the standard clean model set and then to perform further adaptation for the actual test data [7].

Two approaches to initial model derivation have been used: Parallel Model Combination (PMC) and multi-environment training. PMC combines additive and convolutional noise estimates with a model set trained on clean speech to form a more appropriate model for the current conditions. Due to its efficiency, the Log-Add PMC technique [1] was used but it only updates the model mean parameters. Multi-environment training has used a stereo training database containing a clean channel and a "secondary" channel from desk-mounted microphones and then applying single-pass retraining[1] [1].

The later phases of adaptation use the maximum likelihood linear regression (MLLR) technique [4, 5] that we developed for adapting a set of continuous density HMMs. MLLR uses linear transformations of the model parameters (means & variances) to improve the likelihood of the adaptation data. It can use a single global adaptation matrix (single class) or a set of adaptation matrices when more specific transforms are desirable. Although MLLR requires a (word-level) transcription of the adaptation data, since the transformations are very general in nature the technique can be very robust to transcription errors [4]. Furthermore only a small amount of adaptation data is required.

For situations where speech is to be recognised in blocks or sessions that come from the same speaker and environment (as in many transcription tasks), it is possible to perform a number of iterations of MLLR adaptation. On each pass MLLR is used to adapt the models so that they better match the current speaker/environment and then the data is re-recognised with the new model set. This iterative application of MLLR has been found to be most beneficial in circumstances where there is severe mismatch between the data and the initial models. Iterative unsupervised MLLR adaptation can be used to provide both broad global adaptation of the initial models prior to lattice generation and to provide more detailed adaptation in several passes. At each pass the estimate of the speech labels (used for adaptation) is refined and hence more specific MLLR adaptation can be performed.

The technique was developed in the context of the 1995 ARPA H3 task [7] which includes data recorded in unknown noise conditions with unknown microphones. The use of iterative unsupervised adaptation enables high quality word lattices to be generated. Further MLLR passes gradually refine the data transcriptions. For the noisy H3 evaluation data, five such passes were used and the final system gave the lowest error rate in the 1995 H3-P0 evaluation.

The paper is organised as follows. First an overview of the HTK large vocabulary speech recognition system is given and the MLLR technique briefly reviewed. Results are presented on a subset the 1995 H3 development data to illustrate the importance of adaptation before lattice generation. The 1995 H3 evaluation system is then described and the detailed output of each stage of adaptation presented.

[1] The *a posterori* probability of Gaussian occupation is taken from the clean model set/data and then the model parameters are updated using the secondary channel data.

2. HTK RECOGNITION SYSTEM

This section gives an overview of the 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 basic clean speech system can be found in [6].

In the standard system, each speech frame is represented by a 39 dimensional feature vector that consists of 12 mel frequency cepstral coefficients (MFCCs), normalised log energy along with the first and second differentials of these values. Cepstral mean normalisation (CMN) is applied. For the 1995 H3 evaluation system the MFCCs were replaced by a PLP-based [3] cepstral parameterisation and models derived by single-pass retraining.

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 replaced by simple differences.

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 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, with the standard MFCC parameterisation, is the HMM-1 system of [6].

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 [6]) 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.

Gender-independent versions of HMM-1 and HMM-2 were trained on the PLP representation and secondary channel of the SI-284 data. For the clean speech part of the H3 task, gender dependent versions of HMM-2 were also trained on the primary channel of the SI-284 data. Furthermore there were some experiments performed on a PMC-based version of the HMM-1 system.

The HTK LVCSR system uses a time-synchronous decoder employing a dynamically built tree structured network decoder. 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. MLLR OVERVIEW

MLLR was originally developed for speaker adaptation [4, 5] but can equally be applied to situations of environmental mismatch. A set of transformation matrices are estimated which are applied to the Gaussian mean parameters. We have recently extended the approach so that the Gaussian variances can also be updated [2].

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. The mean parameters are usually transformed by a full matrix (in the case of the HTK system a 40×39 matrix) or a block-diagonal matrix which accounts for only the correlations between the statics, 1st differentials and 2nd differentials as appropriate, while the variances are transformed either by a diagonal matrix (as in the experiments here) or by a more complex transform [2].

When only a small amount of data is available, or in cases where very robust transformation estimation is essential, each set of Gaussian parameters (means and variances) are transformed by a single matrix (single regression class case). As more data becomes available , or it is believed that the transcriptions are more reliable, more specific matrices can be computed using only the data that is aligned with that class. In the systems used here, all the speech Gaussians are clustered into a set of 750 base classes, these are then arranged into a hierarchy and the most specific class is generated that has enough observations to robustly estimate the MLLR matrix parameters. The silence models usually form a separate regression class.

MLLR can be applied in a number of different modes including *unsupervised incremental* in which the system generates the labelling and updates the model parameters after every utterance (or after each small block of utterances) and *transcription mode* which processes complete sessions on block (static unsupervised adaptation) as used in this paper.

4. H3 DATA

For the 1995 ARPA H3 task, 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 the development test data the talkers read from US newspaper articles published in 1994 and for the evaluation data the texts were published in 1995.

A subset of the development test data was used for initial experiments and we randomly chose one of the far-field microphones for each speaker. Results are reported on various speaker subsets for the development data. The A-weighted SNR of the development data subset varied from about 11dB to 26dB.

For the evaluation test H3-P0 data one microphone was selected by NIST for each speaker. For the H3-C0 evaluation test the same speech captured by the close-talking microphone was used. 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 using transcription mode adaptation. The A-weighted SNR of the H3-P0 data from each speaker varied from about 7dB to 23dB.

5. LATTICE GENERATION EXPERIMENTS

Many large vocabulary speech recognition systems perform decoding in a number of passes. Word lattices[2] are generated to compactly encode a set of reasonable hypotheses. These lattices are normally produced using simplified acoustic and/or language models. Later decoding stages then use more powerful and complex knowledge sources using the lattice as a word constraint network. If the lattice doesn't contain the correct answer (i.e. lattice word errors occur) then this can markedly reduce the effectiveness of later passes.

The lattices generated by the HTK system contain a set of nodes that correspond to particular time instants and arcs connecting these nodes that represent 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, lattices may contain copies of each word, and further copies can be required to encode the language model constraints.

A number of experiments were performed to assess the different lattice generation strategies on the H3 development test data. In all cases either PMC-based or PLP secondary channel HMM-1 models were used with a 65k word list and a bigram language model trained on 227 million words from the 1994 NAB text corpus.

To evaluate lattice quality the lattice word error rate and lattice density [6] were measured. The lattice error rate gives a lower bound on the word error rate from rescoring the lattice; while the density is the average number of arcs in the lattice per spoken word.

Speaker/ Mic Pair	Baseline % Lattice Error	Prelim 1 % Lattice Error
704/c	26.3	15.5
70a/b	10.3	1.9
70f/d	18.8	7.6
70w/c	24.2	15.8

Table 1: Lattice error rates for several development test speakers with and without a preliminary pass & global MLLR adaptation.

Table 1 shows the lattice word error rates for the four poorest speaker/microphone pairs in our subset of the H3 development test data firstly using the PMC models directly (Baseline) for lattice generation and then generating lattices with models for which a single "preliminary" (Prelim) pass was run using tight beamwidths (to obtain a rough initial transcription) and then models adapted by global MLLR. If clean speech models (rather than PMC-compensated models) had been used directly for lattice generation it is expected that the lattice error rates would several times greater than the PMC lattice error rates.

It can be seen from Table 1 that the use of preliminary MLLR adaptation has reduced the lattice word error rate on average by more than 50%. Another benefit of this approach is that since the lattice generation used more appropriate models the total computational load is reduced.

[2] Some systems use N-Best lists for the same purpose.

Model Set	% Lattice Error	Lattice Density
Baseline	21.2	3650
Prelim 1	11.3	2092
Prelim 2	8.3	1834

Table 2: Lattice quality on sentences from speaker 704/c.

While a single preliminary pass is clearly very effective the transcription used for adaptation is from a poor system. It is expected that superior lattices could be generated with 2 preliminary recognition/MLLR adaptation passes. To examine this possibility the PLP based secondary channel system was used to generate lattices for 10 sentences of speaker 704 (c microphone) and the lattice error rates and lattice densities are given in Table 2 for no adaptation and one and two passes. The table shows that the error rate is further reduced by the second preliminary pass and also that the use of preliminary passes has overall decreased the lattice error rate by 60% while halving the size of the resulting lattices. Hence this approach to lattice generation was adopted for the H3 evaluation.

6. NOV'95 H3 EVALUATION RESULTS

This section describes the the HTK system used for the 1995 H3 evaluation and gives error rates from each of the passes through the data. Unsupervised MLLR is applied for each pass to gradually improve system performance.

6.1. HTK H3 System

The HTK system used for the evaluation test data 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 [6] systems were trained for both paths using the PLP representation. 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. 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. The OOV rate of the test data (accounting for the official mappings used in scoring) was 0.56%.

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 used global MLLR adaptation (i.e. a single transformation for all Gaussians) and the trigram language model. Using the transcrip-

tions from the second preliminary pass, global MLLR adaptation was again performed. These models were used to generate word lattices using a bigram language model. These word lattices had a 3.2% word error rate for the H3-P0 data (and 1.3% for the H3-C0 data) which we believe to be a very significant factor in the good performance of the overall system.

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.

6.2. 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.

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.

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.

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.

7. CONCLUSION

The use of iterative unsupervised MLLR based adaptation has been described in the context of large vocabulary speaker independent continuous speech recognition. It was been shown MLLR adaptation prior to lattice generation can greatly improve lattice quality. The process also improves the speed of lattice generation in such circumstances. The technique can also be applied to gradually improve the data transcription via further recognition (using the lattice constraints) and adaptation passes.

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. Valtcho Valtchev generated the language models used in this work. Additional computing resources were provided by the ARPA CAIP computing facility.

9. REFERENCES

1. Gales M.J.F. (1996). Model-Based Techniques for Noise Robust Speech Recognition. Ph.D. Thesis, Cambridge University.
2. Gales M.J.F., Pye D. & Woodland P.C. (1996). Variance Compensation Within the MLLR Framework for Robust Speech Recognition and Speaker Adaptation. *Proc. ICSLP'96*. Philadelphia.
3. Hermansky H. (1990). Perceptual Linear Prediction (PLP) Analysis for Speech. *J. Acoust. Soc. Amer.*, Vol. 87, pp. 1738-1752.
4. 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.
5. 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.
6. 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.
7. Woodland P.C., Gales M.J.F. & Pye D. (1996). Improving Environmental Robustness in Large Vocabulary Speech Recognition. *Proc. ICASSP'96*, Vol. 1, pp. 65-68, Atlanta.