IMPROVING ENVIRONMENTAL ROBUSTNESS IN LARGE VOCABULARY SPEECH RECOGNITION

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

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

ABSTRACT

This paper describes techniques to improve the robustness of the HTK large vocabulary speech recognition system to non-ideal acoustic environments. The primary methods are single-pass retraining using stereo training data; parallel model combination which combines HMMs trained on clean data with estimates of convolutional and additive noise; and 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 and the 1995 ARPA CSR H3 task (multiple unknown microphones). The HTK system yielded the lowest error rates in both the H3-P0 and H3-C0 tests.

1. INTRODUCTION

Most work on speaker independent large vocabulary continuous speech recognition (LVCSR) has focused on the use of speech recorded using a close-talking noise-cancelling microphone i.e. *clean* speech. Furthermore, the recognition performance of LVCSR systems trained on clean speech and tested in other environments (e.g. different microphones and/or additive noise) tends to be significantly degraded.

This paper investigates model compensation and adaptation techniques to improve the the robustness of the HTK LVCSR system, which gives state-of-the-art performance under clean speech conditions, [8], to non-ideal acoustic environments. It is important that techniques are *data efficient*, i.e. only a small amount of data is required from the new environment to adapt a pre-existing model set. The methods investigated in the following sections include single-pass retraining (SPR) which requires a stereo training database for the new environment (or an approximation to it); parallel model combination (PMC) [1, 2] which combines estimates of convolutional and additive noise to compensate an HMM set trained on clean speech; and 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. Both PMC and MLLR can be viewed as attempting to approximate a SPR-based system in a data efficient manner.

This paper first gives an overview of the HTK LVCSR system and then briefly describes SPR, PMC and MLLR. The performance of these techniques is first evaluated using the 1994 ARPA CSR S5 (alternate microphones) and S10 (additive noise) spoke tasks. Finally the HTK system developed for the 1995 ARPA Hub 3 evaluation is described.

2. CLEAN SPEECH SYSTEM OVERVIEW

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 [8].

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] parameterisation (see Secs. 4 and 5).

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 (the HMM-1 system of [8]) 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 [8]) 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 [7]. 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. TECHNIQUES FOR ENVIRONMENT ADAPTATION

This section describes three techniques for environmental adaptation. Referring to a system trained and tested on data from the same environmental conditions as a *matched* system, all of these techniques change the model parameters to approximate the ideal matched system.

The first method, single-pass retraining, produces a matched system assuming the frame/Gaussian alignment doesn't change between a clean system and a matched system but requires a full stereo training data set. PMC and MLLR need much less data to compensate or adapt the models towards a single-pass retrained system.

3.1. Single-Pass Retraining

Given a mixture Gaussian HMM system trained on clean speech, and assuming that the frame/state (Gaussian mixture component) alignment is identical for any matched system, and that a set of stereo training data exists (or can be synthesised) SPR will produce a matched HMM system.

In a stereo database there are paired speech samples, one clean and the other in the new environment, for the entire HMM training set. In many cases of interest stereo data will be not be available. However if the noise is purely additive and it can be accurately determined, it is feasible to synthesise a stereo training set.

SPR operates by finding the *a posteriori* probability of mixture component occupation using the clean speech models and the clean speech vectors. Once this alignment has been found, the Gaussian parameters are updated using the corresponding observations from the new environment.

SPR provides a baseline against which to measure other approaches which attempt to approximate a matched system. Furthermore, although a stereo database might not exist for the new environment, a *more appropriate* (than clean) stereo data set may exist, and SPR can be used. Other adaptation techniques (such as MLLR) might then be used to model the new environment more closely.

3.2. Parallel Model Combination

PMC attempts to estimate the parameters of an SPR trained matched 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. Furthermore, it is assumed 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.

In the log spectral domain, the ith component of the corrupted speech observation vector, $O_i(\tau)$, is given by

$$O_i(\tau) = \log(\exp(H_i + S_i(\tau)) + \exp(N_i(\tau))) \qquad (1)$$

where H_i is the channel difference between training and test, $S_i(\tau)$ the clean speech and $N_i(\tau)$ the noise at time τ.

Equation (1) can be used to generate "observations" in the new environment which are used to update the static HMM parameters. Similarly "mismatch" functions can be defined for the 1st and 2nd difference parameters and observations generated that include these parameters [1]. Using these new observations, both the mean and variance Gaussian parameters are updated. This version of PMC will be referred to as data-driven PMC or DPMC.

A simpler PMC implementation 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 μ_i is the clean speech mean and $\tilde{\mu}_i$ the noise mean in the log spectral domain. This approximation ensures that compensation is very fast but only the mean parameters can be updated. It will be referred to as Log-Add PMC.

3.3. Maximum Likelihood Linear Regression

MLLR was originally developed for speaker adaptation [5, 6] 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 [3].

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 (in this case 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 normally transformed by a diagonal matrix.

When only a small amount of data is available each set of Gaussian parameters (means and variances) are transformed by a single matrix (single regression class case). As more data becomes available 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 Gaussians for 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).

4. 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.

To save computation, some results were generated using pre-computed word lattices. Since the lattices were generated by systems that were well-matched to the test-data, optimistic results may be generated for poor systems. In the tables results in parentheses were generated with intermediate lattices from a different system.

In all PMC experiments, the channel mismatch, H_i, was estimated in the manner described in [2] using a 30-component Gaussian mixture model and the first sentence from each speaker.

4.1. 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. Since the test data is only corrupted by additive noise and a noise sample was available, it is possible to synthesise a stereo database for SPR to evaluate the PMC and MLLR approaches. PMC used a noise model built using the background noise sample provided with the dataset.

Table 1 shows that performance with the clean models is very poor: the error rate on the corresponding clean data with the standard MFCC parameterisation is 5.8% and 6.7% with that used with PMC. SPR is very effective in reducing the error rate and updating just the Gaussian means (shown in Table 1 as SPR (Mean)) is nearly as effective as updating both the means and variances. The MLLR (Class) results give the error rates using multiple MLLR regression matrices and different amounts of data while the MLLR (Global) uses a single regression matrix for all data. The MLLR systems were trained on subsets of the stereo SPR data. MLLR is effective in producing models that approximate the SPR models with much less data.

Training Type	Adaptation Sentences	% Word Error Rate
Clean	—	54.3 (33.9)
SPR	36493	10.1 (9.9)
SPR (Mean)		10.5 (9.9)
MLLR (Class)	2000	(10.5)
	200	(10.6)
	40	(11.0)
MLLR (Global)	40	(12.1)
DPMC	—	10.3
DPMC (Mean)		10.7
PMC Log-Add		10.7 (9.6)

Table 1. Performance on S10 Level 3 data.

PMC gives similar results to MLLR, but has the advantage of not needing any stereo training data to be synthesised. Again there is only a small degradation for updating only the mean parameters. Note that even in the case of purely additive noise, an H_i term is estimated that performs gain matching and simple speaker adaptation.

Updating Means	Updating Variances	Word Error (%)
$\times$	$\times$	10.7
$\surd$	$\times$	9.3
$\surd$	$\surd$	8.9

Table 2. Unsupervised incremental MLLR on PMC Log-Add S10 system.

Table 2 shows how incremental adaptation on the PMC Log-Add system further reduces the error rate by about 17% with variance adaptation contributing 4%.

4.2. S5 Experiments

The S5 data consisted of 200 sentences from 20 speakers. For each speaker one of 10 alternate microphones was used. The A-weighted SNR was typically 20dB.

In this case it isn't possible to produce a paired stereo training dataset (since the channel effect and additive noise is unknown) but results from a roughly matched system trained on the "secondary channel" data of the SI-284 training set are reported. This secondary channel data 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 HMM-1 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	10.6	8.6	8.0
PLP 2nd channel	9.0	7.4	7.1

Table 3. % Word error rates for S5 data.

Table 3 shows that 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. Incremental MLLR adaptation again provides improvements with variance compensation further decreasing the error rate by about 5%.

5. NOV'95 H3 EVALUATION SYSTEM

5.1. 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.

5.2. 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 [8] 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. 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 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 [7] 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.

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 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.

5.3. H3 Results

The last line of Table 4 gives the actual HTK results in the Nov'95 H3 evaluation. These were the lowest error rates achieved in both the H3-P0 and H3-C0 tests. The results use the adjudicated transcriptions and map files.

The other lines of Table 4 show the result of using the HMM-2 models with the 4-gram lattices derived above 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

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. † denotes the systems actually used for the Nov'95 H3 evaluation.

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.

6. CONCLUSION

The techniques described, used both singly and together, have been shown to produce a large decrease in word error rate when a LVCSR system is used in environmental acoustic conditions that aren't matched to the training data. The resulting systems give state-of-the-art performance on data with both additive noise and channel effects.

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 built the language models for the H3 system. Additional computing resources were provided by the ARPA CAIP computing facility.

REFERENCES

[1] Gales M.J.F. & Young S.J. (1995). A Fast and Flexible Implementation of Parallel Model Combination. *Proc. ICASSP'95*, Vol. 1, pp. 133-136, Detroit.

[2] 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.

[3] Gales M.J.F. & Woodland P.C. (1995). 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] 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.

[8] Woodland P.C., Leggetter C.J., Odell J.J., Valtchev V. & Young S.J. (1994). The 1994 HTK Large Vocabulary Speech Recognition System. *Proc. ICASSP'95*, Vol. 1, pp. 73-76, Detroit.