# EXPERIMENTS IN SPEAKER NORMALISATION AND ADAPTATION FOR LARGE VOCABULARY SPEECH RECOGNITION

*D. Pye & P.C. Woodland*
Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, England.

## ABSTRACT

This paper examines techniques for speaker normalisation and adaptation that are applied in training with the aim of removing some of the variability from the speaker independent models. Two techniques are examined: vocal tract normalisation (VTN) which estimates a single "vocal tract length" parameter for each speaker and then modifies the speech parameterisation accordingly and speaker adaptive training (SAT) which estimates Gaussian mean and variance parameters jointly with a speaker specific set of maximum likelihood linear regression (MLLR) based transformations. It is shown that VTN is effective for both clean speech and mismatched conditions and that the further improvements obtained by applying MLLR in testing are essentially additive. Detailed results from the use of SAT show that worthwhile improvements over using MLLR with standard speaker independent models are obtained.

## 1. INTRODUCTION

Recently there has been much interest in adaptation techniques for large vocabulary speech recognition. Normally adaptation is applied to a baseline speaker independent recognition system that has been trained by pooling all the training data and modelling the variations found in the pooled data. For a system that will, in use, be adapted to a particular speaker (or environment) a better initial system can be constructed by taking into account the type of adaptation that will be applied in testing and training models accordingly. The resulting models are, at least in part, *normalised* to speaker variability. The simplest type of normalised training, which is now in widespread use, is to use cepstral mean normalisation (CMN). In its commonest form CMN removes the mean cepstrum from all vectors with the cepstral mean calculated separately for each sentence. In this paper two more complex techniques for using adaptation in training are discussed: vocal tract normalisation (VTN) and speaker adaptive training (SAT).

VTN [2, 5] normalises the data by using a linear frequency scaling to try to account for the effect of the variations in vocal tract length on formant frequencies. The vocal tract length scaling factors are estimated by performing a search over a set of possible factors and choosing the one which maximises the data likelihood based on a set of HMMs. Since VTN only requires a single parameter to be estimated it can be applied on an utterance by utterance basis during recognition.

One topic of particular interest is whether the improvements from VTN are complementary to those obtained by maximum likelihood linear regression (MLLR) [6, 7, 4] based adaptation. MLLR estimates a set of linear transformations for the Gaussian parameters to maximise the likelihood of adaptation data. Although MLLR is more complex than VTN in terms of numbers of parameters, neither technique can directly account for the same effects as the other.

SAT [1] estimates the parameters of the Gaussian means and variances assuming that the Gaussian means have been transformed by MLLR. In training the standard re-estimation formulae are modified so that the MLLR based transforms, the Gaussian means and the variances are jointly estimated in an iterative process.

In this paper an experimental evaluation of VTN and SAT is performed using the HTK large vocabulary system which gives state-of-the-art performance in both matched and mismatched conditions [9] [10]. The next sections describe our approach to VTN and give experimental results for using VTN with and without MLLR in both clean and mismatched environments. A brief description of the SAT technique is then given and recognition results at different stages of training are presented. It is shown that both types of adaptation in training provide useful improvements in performance.

## 2. VOCAL TRACT NORMALISATION

The primary effect of vocal-tract length variation between speakers is a linear scaling of frequency. The aim of vocal tract length normalisation is to estimate a length (or frequency) scaling factor for each speaker (or utterance) and then normalise the speech signal to an average vocal tract length so the parameterised speech is independent of this type of inter-speaker variation.

Recent interest in this approach stems from the work at the 1994 CAIP Summer Workshop on the Switchboard telephone corpus [2]. This work estimated a vocal tract length factor for each speaker and then resampled the speech waveform to effect the linear frequency scaling. Other work on this topic has also concentrated on the Switchboard corpus [3] or other telephone corpora [5] [8] whereas in this paper we examine the use of VTN for wide-bandwidth corpora with both known and unknown microphones.

The main issues to be addressed in an implementation of VTN are the estimation of the scaling or frequency warping factors and the implementation of the frequency scaling in the speech parameterisation. The frequency scaling can be estimated either by searching a discrete set of possible scalings [2] or using a more direct approach based on measuring e.g. formant frequencies [3]. For the methods based on search, the utterance is processed a number of times for each putative warp frequency and the maximum likelihood

To appear in *Proc. ICASSP-97, Apr 21-24, Munich, Germany*

warp factor given a set of speech HMMs is chosen. Frequency scaling can either be implemented by resampling in the time domain [2], or more efficiently, for filter-bank based front-end processing, by modifying the filter-bank centre frequencies for each warp factor [5].

In the work here we have adopted a search approach and filter-bank based normalisation. Following [5] 13 vocal-tract length scaling factors are considered in steps of 0.02 from 0.88 to 1.12 and the speech data coded for each of these different scalings by adjusting the filter-bank centre frequencies. In such an implementation care must be taken to deal with the highest frequency channel when the centre frequencies are increased beyond the range of the usual analysis. Here, in such cases we have estimated the highest channel energy by interpolating with the neighbouring channels and in extreme cases (length warpings of 0.88 or below) copied the highest channel energy from the next highest frequency bin.

In model training a single warp frequency is selected for each speaker by examining the likelihood of the warped data at each scaling using a set of HMMs corresponding to the known transcriptions. It may be beneficial to iteratively estimate the training data frequency warpings since the initial selection of scalings is based on models trained from unwarped data.

In recognition ideally a full recognition pass with the data at each scaling would be performed and the scaling that gave the maximum likelihood output chosen. However this procedure is computationally very expensive so instead the likelihood of the data using the transcription from a first pass decoding with unwarped data is used to select the frequency scaling. The selected scaling factor is then used and the data re-recognised (possibly using lattices from the first pass) to generate the final recognition output.

## 3. VTN EXPERIMENTS

There were two main motivations of these experiments: to investigate the performance of VTN on large vocabulary non-telephone data and to establish whether gains from VTN and subsequent MLLR adaptation are additive.

The VTN experiments employ the HTK LVCSR system [9] which uses state-clustered, cross-word triphone mixture Gaussian HMMs. The configuration used here has an MFCC front-end supplemented with 1st and 2nd differentials and gender independent models. Two different sets of experiments were performed: the first used just the WSJ0 SI-84 training data (14 hours of speech) while the second set used the full WSJ0+1 SI-284 training set (66 hours).

### 3.1. SI-84 Models

The experiments with SI-84 models used a model set containing 3992 clustered speech states and 8 Gaussian components per state. The experiments used two test data sets: the US SQALE evaluation data and the 1994 ARPA S5 data. The SQALE data consists of 200 WSJ utterances from 20 speakers recorded in clean conditions. The ARPA November 1993 20k trigram language model and word list were used with the SQALE data. The S5 data is used to test performance for non-matched microphones and consists of 200 sentences from 20 speakers. For each speaker one of 10 alternate microphones was used. The A-weighted SNR was typically 20dB. For S5 the standard ARPA 5k trigram was used.

The results of the experiments on both the S5 and SQALE test-sets are shown in Table 1. The first line of the table gives the non-normalised training and test baseline system results; then the result of normalising the test data only and finally normalising both training and test with either one or two iterations of normalisation. The Table shows that VTN gives a 15% reduction in word error rate on S5 and 10% for the SQALE data. We have previously found on the SQALE data that the use of gender dependent modelling gives just a 2% improvement in error rate so VTN is clearly much more effective. It was also noted that all the normalisation factors for S5 tend to have a compressive frequency scaling showing that some environmental normalisation is also occurring for this data.

| Test Data Normalised | Training Data Normalised | % Word Error Rate | |
|---|---|---|---|
| | | S5 | SQALE |
| N | N | 13.9 | 13.6 |
| Y | N | 13.3 | 12.7 |
| Y | 1 Iteration | 12.0 | 12.1 |
| Y | 2 Iterations | 11.8 | 12.2 |

**Table 1. VTN results on the S5 and SQALE data**

We next investigated the effect of combining VTN and MLLR for the S5 and SQALE data. A model set trained using normalised utterances was transformed by MLLR to maximise the likelihood of the warped test utterances. MLLR was used in transcription mode on the test data: the test data is used as the adaptation data with the recognised output as the transcription; the model parameters are transformed and the data is then re-recognised. Two such passes of MLLR were performed and each MLLR pass updates the means using block-diagonal transforms and the variances with a diagonal transform (plus offset). A global transformation was used in a first MLLR pass and then, in the second pass, multiple classes determined by a regression class tree [7] were used. For both passes a separate silence class was used.

| Task | Speaker Normalised | % Word Error Rate | | |
|---|---|---|---|---|
| | | No MLLR | Global | 2nd Pass |
| S5 | — | 13.9 | 9.0 | 8.4 |
| S5 | VTN | 11.8 | 8.4 | 7.5 |
| SQALE | — | 13.6 | 12.1 | 11.9 |
| SQALE | VTN | 12.2 | 10.7 | 10.6 |

**Table 2. VTN with MLLR results on S5 and SQALE data**

Table 2 shows the results of applying MLLR to the baseline and the best VTN system on both data sets. On the clean SQALE data, the gains from both VTN and MLLR are clearly additive with a 10-12% error reduction due to VTN and 13% due to MLLR. We found this to be a somewhat surprising but very encouraging result. For the S5 data, MLLR provides a 40% reduction in error rate since both environmental compensation and speaker adaptation are being performed whereas VTN alone gives a 15% reduction in error rate and a further 11% when combined with MLLR, showing that the gains are still largely additive.

### 3.2. SI-284 Models

Most speaker adaptation techniques become relatively less effective when applied to more sophisticated initial models trained on larger training corpora. It was expected that this would be the case for VTN and hence the above experiments were repeated using HMMs trained on the SI-284 (WSJ0+1) training data.

The models used were gender independent cross word state clustered triphones with 6399 speech states and 12 mixture components per state (the HMM-1 set of [9]). In order to reduce time in developing a VTN compensated model set, a single warping factor was estimated for each speaker by averaging the warping factors for a subset of 20 utterances from each speaker.

Recognition used a 65k trigram language model and operated in lattice rescoring mode. The test data consisted of the 1994 ARPA H1 development and evaluation data (unlimited vocabulary data). Each of these data sets contains approximately 15 sentences from each of 20 speakers.

| Test Data Normalised | Training Data Normalised | % Word Error Rate | |
|---|---|---|---|
| | | H1 Dev | H1 Eval |
| N | N | 9.35 | 9.07 |
| Y | N | 8.92 | 8.61 |
| Y | Y | 8.63 | 8.43 |

**Table 3. VTN results on the H1 Dev and Eval tasks.**

The results for Table 3 shows VTN performance on these tasks. The use of test-only VTN reduces the error rate by 5%, while using normalised data for both training and test reduces the error rate by 7-8% i.e. a little less than the 10% reduction in error observed earlier on clean data using the SI-84 models. We conclude that although VTN is less effective as more training material is available, it is still beneficial to systems trained on large quantities of data.

| Task | Speaker Normalised | % Word Error Rate | | |
|---|---|---|---|---|
| | | No MLLR | Global | 2nd Pass |
| H1 Dev | — | 9.35 | 8.45 | 8.23 |
| H1 Dev | VTN | 8.63 | 7.81 | 7.70 |
| H1 Eval | — | 9.07 | 8.11 | 7.83 |
| H1 Eval | VTN | 8.43 | 7.47 | 7.32 |

**Table 4. VTN with MLLR results on H1 Dev and Eval data.**

In Table 4, the effects of MLLR on the baseline and VTN SI-284 model set is shown. MLLR has been used as described earlier: an initial global pass is followed by a multiple class transformation using transcription mode adaptation on the test data. On the H1 Dev set an 8% error rate reduction due to VTN alone is observed and 12% due to MLLR alone, with an 18% improvement when both are used. Similarly on H1 Eval, a gain of 7% from VTN alone and 14% using MLLR alone compares to a combined gain of 19%. Therefore although the improvements due to VTN are a little smaller with a larger model set, as with the SQALE task described earlier using SI-84 models, the improvements due to MLLR and VTN again appear to be additive.

## 4. SPEAKER ADAPTIVE TRAINING

MLLR adaptation is normally used with a standard speaker independent model set created by simply pooling all the training data rather than with models specifically created for adaptation. However, if it is known that the model parameters will be transformed using speaker-specific MLLR a more appropriate set of initial mean and variance parameters can be found. This is the basis of speaker adaptive training (SAT) [1]. The basic idea is that the inter-speaker variability that can be accounted for by speaker-specific MLLR is in effect removed when estimating the mean and variance parameters. It may be expected that a smaller number of Gaussians is needed than the standard SI training approach to model the data with the same detail.

For SAT training it is necessary to jointly estimate a set of transformation matrices for each of the training speakers (which depends on the mean and variances of the models) and values for the Gaussian means and variances (which depend on the transformations). Following [1], an iterative approach is adopted in which one of these parameter sets (transformations, means, variances) is estimated at each stage and maximum likelihood re-estimation used individually for each of the parameter sets assuming the other parameters are fixed. Once all the parameters have been updated in this way, further complete iterations are possible.

The transformation matrices are estimated using the standard MLLR equations; the model means are updated given that the means are transformed by speaker-dependent MLLR transformations; and finally the variances are found given the means and the transformations. The formula for the mean update of a particular Gaussian $m$ is given by (assuming a single training utterance for each of $S$ speakers for notional convenience) [1]

$$\hat{\mu}_m = \left( \sum_{s=1}^{S} \sum_{\tau=1}^{T_s} L_m^s(\tau) \mathbf{A}_s^T \mathbf{\Sigma}_m^{-1} \mathbf{A}_s \right)^{-1} \times \sum_{s=1}^{S} \sum_{\tau=1}^{T_s} L_m^s(\tau) \mathbf{A}_s^T \mathbf{\Sigma}_m^{-1} \left( \mathbf{o}^s(\tau) - \mathbf{b}_s \right)$$

where $L_m^s(\tau)$ is the *a posteriori* probability of Gaussian $m$ at time $\tau$ for speaker $s$; $T_s$ is the length of the utterance from speaker $s$; $\mathbf{o}^s(\tau)$ is an observation at time $\tau$; $A_s$ is a square MLLR transformation matrix for speaker $s$ and Gaussian $m$ and $\mathbf{b}_s$ is the corresponding MLLR offset vector.

The variance update is given by [1]

$$\hat{\mathbf{\Sigma}}_m = \frac{\sum_{s=1}^{S} \sum_{\tau=1}^{T_s} L_m^s(\tau)(\mathbf{o}^s(\tau) - \mu_m^s)(\mathbf{o}^s(\tau) - \mu_m^s)^T}{\sum_{s=1}^{S} \sum_{\tau=1}^{T_s} L_m^s(\tau)}$$

where

$$\mu_m^s = \mathbf{A}_s \mu_m + \mathbf{b}_s$$

It should be noted that the mean update in particular is considerably more computationally intensive than standard maximum likelihood training, and in particular a straightforward implementation requires significantly increased storage for the "denominator" of the mean update.

Experiments using SAT training with this 3-step optimisation of transformations means and variances were reported in [1] for a phonetically tied mixture HMM system with a relatively small number of Gaussians.

## 5. SAT EXPERIMENTS

The aim of these experiments was to evaluate the SAT technique using a state-clustered Gaussian mixture system and to discover the contribution of training the mean and variance parameters using SAT.

The HMM system used an MFCC front-end and was trained using the SI-284 WSJ0+1 data. The system was initialised using a set of standard speaker independent models. These were gender independent cross word state clustered triphones with 6399 speech states and 12 mixture components per state. This is the same as the baseline system described in the VTN section for experiments on the ARPA 1994 H1 development and evaluation tasks [9]. These test sets were again used for SAT experiments using a 65k trigram language model in lattice rescoring mode.

In training, block-diagonal transformations were used and were trained from all the training data for that speaker in the SI-284 set. A regression-class tree was used to define the sets of transformations trained and thresholds were set in transform creation so as to ensure that a similar number of transformations were used in both the training and test phases. The model parameters were trained for 2 complete iterations of the 3-step optimisation starting from the baseline standard speaker independent system. For each iteration first training transforms were estimated (a); then the mean parameters were updated (am) and finally the variances (amv).

In testing, the 40 standard adaptation sentences were used for each speaker in static supervised adaptation mode to train a set of block-diagonal mean transformations using a regression class tree. To obtain accurate state-frame alignments for transformation estimation the test adaptation matrices were estimated for each stage of model training in turn.

It should be noted that the experiments in this section are not directly comparable to the VTN experiments since different adaptation data (& adaptation mode) is used and also MLLR is used to update the mean parameters only.

| Model Set | Test MLLR | % Word Error Rate | |
|---|---|---|---|
| | | H1 Dev | H1 Eval |
| Baseline | N | 9.35 | 9.07 |
| Baseline | Y | 8.03 | 8.13 |
| am | Y | 7.90 | 7.85 |
| amv | Y | 7.85 | 7.55 |
| amvam | Y | 7.69 | 7.43 |
| amvamv | Y | 7.73 | 7.37 |

**Table 5. SAT results on the 1994 H1 Dev and Eval data**

Table 5 shows the baseline error rates and the performance at each stage of SAT training for both the H1 development and H1 evaluation data. Standard MLLR reduces the error rate by an average of 12%, and SAT based training reduces the error rate by a further 6% with the second iteration of SAT training (amvam and amvamv sets) providing about 1/3 of this reduction. It can also be seen that the adjustment of both the means and the variances contributes to the overall improvement obtained.

In the experiments here SAT has been found to give worthwhile improvements in word error rates and the resulting models certainly have a more speaker dependent character than standard mean-only MLLR-adapted models. In particular due to the reduced variances of the SAT trained models recognition speed is noticeably improved since pruning in decoding is more effective.

Although all the SAT experiments here have used the same number of parameters as the standard speaker independent system, it would be possible to train a system completely using the SAT approach so that the final system has a reduced number of parameters. Such a system, combined with static supervised adaptation, could be a particularly effective way of training compact speaker-dependent models from a limited amount of enrolment data.

## 6. CONCLUSIONS

This paper has investigated the use of both vocal tract normalisation and speaker adaptive training techniques. Both methods are shown to be effective since they are able to effectively reduce the variability that needs to be modelled. One interesting result is that the use of VTN is shown to be essentially additive to gains from standard MLLR adaptation. Further work may include examining these two techniques in combination.


## ACKNOWLEDGMENTS

This work is in part supported by an EPSRC grant reference GR/K25380. Mark Gales provided helpful advice concerning SAT training. The SAT experiments were performed whilst PCW was visiting CNRS-LIMSI in Paris. The support of LIMSI and the Université de Paris Sud is gratefully acknowledged.


## REFERENCES


[1] Anastasakos T., McDonough J., Schwartz R. & Makhoul J. (1996) A Compact Model for Speaker Adaptive Training. *Proc. ICSLP'96*, pp. 1137-1140, Philadelphia.

[2] Andreou A., Kamm T. & Cohen J. (1994). Experiments in Vocal Tract Normalisation. *Proc. CAIP Workshop: Frontiers in Speech Recognition II.*

[3] Eide E. & Gish H. (1996). A Parametric Approach to Vocal Tract Length Normalisation. *Proc. ICASSP'96*, Vol. 1, pp. 346-348, Atlanta.

[4] Gales M.J.F. & Woodland P.C. (1996). Mean and Variance Adaptation Within the MLLR Framework. *Computer Speech & Language*, Vol. 10, pp. 249-264.

[5] Lee L. & Rose R.C. (1996). Speaker Normalisation Using Efficient Frequency Warping Procedures. *Proc. ICASSP'96*, Vol. 1, pp. 353-356, Atlanta.

[6] 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.

[7] Leggetter C.J. & Woodland P.C. (1995). Flexible Speaker Adaptation For Large Vocabulary Speech Recognition. *Proc. Eurospeech'95*, Vol. 2, pp. 1155-1158, Madrid.

[8] S. Wegmann, D. McAllaster, J. Orloff, B. Peskin (1996). Speaker Normalisation on Conversational Telephone Speech. *Proc. ICASSP'96*, Vol. 1, pp. 339-341, Atlanta.

[9] 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.

[10] 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.