INVESTIGATION OF ACOUSTIC MODELLING TECHNIQUES FOR LVCSR SYSTEMS

M. J. F. Gales, X. Liu, K. C. Sim & K. Yu

Cambridge University Engineering Dept,
Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {mjfg,xl207,kcs23,ky219}@eng.cam.ac.uk

ABSTRACT

The CUHTK evaluation systems typically make use of a multiple pass, multiple branch, framework. This allows a range of acoustic models to be used in the framework and the output from all the systems, or branch, to be combined to give the final output. This paper describes experiments with several advanced acoustic modelling techniques that were candidate approaches for the 2004 CU-HTK large vocabulary speech recognition systems. These techniques include Gaussianization for speaker normalization, discriminative cluster adaptive training, discriminative subspace for precision and mean modelling of inverse covariances, and discriminative complexity control. Acoustic models built using these techniques were integrated into a state-of-the-art 10 real-time multi-pass system with sophisticated adaptation for performance evaluation. Experimental results are presented on both broadcast news (BN) and conversational telephone speech (CTS) transcription tasks.

1. INTRODUCTION

For many years automatic transcription of broadcast news (BN) and conversational telephone speech (CTS) data have been the two main tasks for the research community of large vocabulary continuous speech recognition (LVCSR). Due to the difficulty of these tasks, a variety of modelling techniques have been developed to allow systems to model highly complex data and be robust to changes in acoustic environment. In this paper several advanced modelling techniques that were candidates techniques for the CUHTK 2004 BN-English and CTS-English evaluation systems. The approaches are investigated in the framework of a state-of-the-art multi-pass LVCSR system using sophisticated adaptation, large scale language models and Confusion Network (CN) based system combination. By implementing the approaches in this complex framework, it is possible to obtain a realistic estimate of how they may perform in an evaluation style system. Techniques investigated include Gaussianization for speaker normalization, discriminative Cluster Adaptive Training (CAT), Subspace for Precision And Mean (SPAM) modelling of inverse covariances, and model complexity control.

Some of these approaches investigated, such as SPAM, yield systems which do no have diagonal covariance matrices. It is therefore not possible to use the standard efficient techniques for estimating the adaptation transformations in schemes such as Maximum Likelihood Linear Regression (MLLR). This paper also addresses this problem. Rather than using generic gradient descent style optimization [1], simple iterative schemes closely related to the standard optimization approaches are described, along with simpler approximate schemes.

The rest of the paper is organized as follows. Section 2 describes the four acoustic modelling techniques examined. Section 3 examines the adaptation schemes used. In particular, the efficient row-by-row update approach for MLLR mean and constrained MLLR (CMLLR) adaptations of the SPAM model are described. Section 4 gives an overview of the basic features of the CU-HTK 10xRT system. Experimental results are given for various adaptation configurations for the SPAM models. Then experimental results of individual and combined systems on both BN and CTS transcription tasks are presented. Section 5 is the conclusion.

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred. The authors would like to thank G. Evermann for the initial setup of the 10xRT framework and scripts for system evaluations.

2. MODELLING TECHNIQUES

This section describes the theory of Gaussianization, CAT, SPAM and discriminative complexity control. Some implementation issues are also discussed for individual techniques.

2.1. Gaussianization

Cepstral mean and variance normalization is a simple speaker normalization scheme. The aim is to transform the distribution of a speaker's data to distribution having zero mean and unit variance. However, the approach does not attempt to normalize the higher-order moments of the distribution. For scenarios where there is highly non-homogeneous speech data, such as broadcast news, additional normalization of the higher-order moments may be beneficial. In this paper a non-linear speaker normalization scheme, *Gaussianization*, is investigated for both BN and CTS tasks. The idea of Gaussianization is to transform the distribution of an individual speaker to be a standard Gaussian. The approach adopted is to separately model each dimension of a speaker by a on-dimensional Gaussian Mixture Model (GMM). Using the Cumulative Density Function (CDF) of this GMM, it is possible to transform any observation so that the overall distribution for that dimension is a normal Gaussian. Note this does not guarantee that the distribution for the complete feature vector is a normal multivariate Gaussian. The approach is similar to the one described in [2]. However rather than using histogram normalization, a GMM is used to model the data. It is also related to the Gaussianization scheme described in [3], though iterative Gaussianization is not performed.

Let o_j denote the j^{th} dimension of a n dimensional acoustic feature vector o of speaker s. Then the Gaussianized feature on j the dimension is given by,

$$\breve{o}_j^{(s)} = \phi^{-1} \left(\int_{-\infty}^{o_j^{(s)}} \sum_{m=1}^{M_{sj}} c_{sjm} \mathcal{N}\left(x;\, \mu^{(sjm)}, \sigma^{(sjm)2}\right) \mathrm{d}x \right) \quad (1)$$

where $\phi^{-1}(\cdot)$ denotes the standard Gaussian inverse CDF. The speaker GMM component mean, variance and prior is denoted by $\mu^{(sjm)}$, $\sigma^{(sjm)2}$ and c_{sjm} respectively. For each speaker a total of n single dimension M_{sj} component GMMs are trained using Maximum Likelihood (ML) criterion. This scheme provides a more compact and smooth representation of the target distribution than the histogram scheme in [2].

In this work Gaussianization was performed on top of HLDA projected cepstral features. The normalized features were then used in both training and testing. All GMMs used for Gaussianization had 4 components.

2.2. Cluster Adaptive Training

Multiple-cluster schemes, such as cluster adaptive training (CAT) or eigenvoices system, are popular approaches for rapid speaker and environment adaptation [4]. Here, a multiple-cluster model is used as the canonical model in an adaptive training framework. A set of interpolation weights are used to transform this multiple-cluster model to a standard HMM set representative of an individual speaker or acoustic environment which is then used in decoding. Usually only multiple-cluster means are considered. Thus adapted mean vector is represented as

$$\boldsymbol{\mu}^{(sm)} \;=\; \mathbf{M}^{(m)} \boldsymbol{\lambda}^{(s)} \quad (2)$$

where $\mathbf{M}^{(m)} = [\boldsymbol{\mu}_1^{(m)}, \ldots, \boldsymbol{\mu}_P^{(m)}]$ is the multiple-cluster mean matrix, $\boldsymbol{\lambda}^{(s)} = [\lambda_1^{(s)}, \ldots, \lambda_P^{(s)}]'$ is the interpolation weight vector.

Maximum likelihood estimation for the multiple-cluster model and interpolation weights are described in [4]. Initializations of CAT is also detailed discussed in the paper, which allows CAT to be used in LVCSR systems. However, to get state-of-the-art performance, discriminative training, particularly minimum phone error (MPE) training is required. This has been studied for multiple cluster systems in [5]. Though both model parameters and interpolation weights can be discriminatively updated, a simplified version of discriminative adaptive training is commonly used, in which ML-estimated weights are fixed in later discriminative training stage.

In the CU-HTK 10xRT system, the estimation of the test-set transformations for the CAT system employed a similar approach to the standard SAT system. CAT weight transforms are iteratively estimated using the ML criterion based on supervision from the previous lattice generation stage. Then given using these transforms, the standard adaptations transforms can be estimated in a cascade fashion for lattice rescoring using the standard CAT adapted models. More details of the evaluation framework are given in section 4.1.

2.3. Precision Matrix Modelling

The most commonly used form of GMMs for speech recognition use diagonal covariance matrices. Structured precision matrix approximations have been found to yield improved performance using both ML and MPE training [1, 6]. They yield a compact representation and efficient likelihood calculation. Examples of this form of model are the Semi-tied Covariances (STC), Extended Maximum Likelihood Linear Transform (EMLLT) and Subspace for Precision And Mean (SPAM) systems. The precision matrix (inverse covariance), $\boldsymbol{P}_m$, of a Gaussian component m, can be expressed in a general form of basis superposition:

$$\boldsymbol{P}_m \;=\; \sum_{i=1}^{n} \lambda_{ii}^{(m)} \boldsymbol{S}_i \;=\; \sum_{i=1}^{n} \lambda_{ii}^{(m)} \sum_{r=1}^{R} \lambda_{rr} \boldsymbol{a}_{ir}' \boldsymbol{a}_{ir} \quad (3)$$

where $\boldsymbol{S}_i$ is the ith basis matrix and $\lambda_{ii}^{(m)}$ is the corresponding basis coefficient. $\boldsymbol{S}_i$ is a symmetric matrix with an arbitrary rank, R, which can be further decomposed into a superposition of R basis vectors, $\boldsymbol{a}_{ir}$. $\boldsymbol{P}_m$ is constrained to be positive-definite. If, $\boldsymbol{S}_i$ is rank-1 ($R = 1$), equation 3 becomes a STC model when $n = d$ and an EMLLT model when $d < n \leq \frac{d}{2}(d+1)$. Removing the rank-1 constraint gives the SPAM model. Among these, SPAM was found to yield the best performance [6]. This paper will also consider SPAM modelling within a speaker adaptively trained feature space.

This paper considers MPE discriminatively trained SPAM models. Two variants of SPAM models were trained. The first model was trained within the 39-dimensional HLDA feature space. The second form of model was built with an adaptively trained feature-space. Here constrained MLLR was used to generate a standard ML Speaker Adaptively Trained (SAT) system. Then within the adaptively trained feature-space the precision matrix models were built. This is the SAT-SPAM system

2.4. Complexity Control

There are a wide range of possible models that can be used for LVCSR. It is not practical to build, and compare, each possible system for large vocabulary speech recognition systems. To overcome this problem automatic model complexity control schemes have been proposed [7, 8]. Most existing complexity control schemes make an assumption that increasing the likelihood on held-out data can decrease the word error rate (WER). However this correlation has been found quite weak for current speech recognition systems. It would be preferable to use a criterion more closely related to WER. One possible method is to marginalize a discriminative criterion. However, due to sensitivity to outliers, discriminative training criteria, such as Maximum Mutual Information (MMI), can not be directly integrated for complexity control.

To overcome this problem the marginalization of a discriminative growth function has been proposed [9]. The growth function of a discriminative criterion retains its curvature in the parametric space, and largely removes the sensitivity to outliers. Let λ denotes the model parameters. For a family of discriminative criteria that can be expressed as a ratio between two polynomials with positive coefficients (including MMI and MPE), $\mathcal{F}(\lambda) = \mathcal{F}_{\mathrm{num}}(\lambda)/\mathcal{F}_{\mathrm{den}}(\lambda)$, a generic form of the associated growth function is given below.

$$\mathcal{G}(\lambda) \;=\; \mathcal{F}_{\mathrm{den}}(\lambda) \left[\mathcal{F}(\lambda) - \mathcal{F}(\tilde{\lambda}) + C \mathcal{F}_{\mathrm{sm}}(\lambda, \tilde{\lambda}) \right] \quad (4)$$

where $\tilde{\lambda}$ is the *current* parameter estimate. The first two terms in the bracket retain the criterion's curvature in the parametric space. A third smoothing criterion or statistics, $\mathcal{F}_{\mathrm{sm}}(\lambda, \tilde{\lambda})$, scaled by a constant $C > 0$, acts to remove the sensitivity to outliers by by penalizing highly unlikely word sequences. The exact form of the smoothing term depends on the underlying discriminative criterion being considered. Using a generalized EM approach, a strict lower bound of the growth function can be derived. This has a more tractable form for marginalization, with the dependence on the hidden variables removed. A second order Laplace's approximation can be used for the growth function integration.

In this paper complexity controlled acoustic models were built using this marginalized growth function. Two forms of complexity were varied. In contrast to the standard global 39-dimension HLDA projection, the systems were built with multiple HLDA transforms, in this case 65, with number of retained dimensions varied. In addition the number of components per state were varied. Both forms were determined using a marginalized MPE criterion. For more details of the experimental set-up see section 4.1.

3. ADAPTATION OF ACOUSTIC MODELS

An important aspect of any form of improved acoustic models is the applicability of adaptation techniques to these models. This paper considers three forms of the MLLR adaptation schemes, namely the mean [10], covariance and constrained MLLR (CMLLR) [11] adaptation schemes. Gaussianization is a feature transformation scheme. So, the form of adaptation methods required only depends on underlying model used to represent the Gaussianized features. In this paper, Gaussianization systems are built using the diagonal covariance matrix models where the standard adaptation approaches given in [10, 11] can be directly employed. The following sections will describe the application of the above mentioned adaptation schemes to CAT, SPAM and complexity control systems. Since the HLDA system can be viewed as a special form of basis superposition precision matrix model with basis coefficient tying for the *nuisance* parameters [12], the multiple HLDA projections used in the complexity control system is equivalent to a multiple basis precision matrix model. Hence, the following adaptation schemes will be described in terms of a generic form of precision matrix structure.

3.1. MLLR Mean Adaptation

MLLR adaptation of the mean vector [10] can be written as

$$\hat{\boldsymbol{\mu}}_m = \boldsymbol{A}^r\boldsymbol{\mu}_m + \boldsymbol{b}^r = \boldsymbol{W}^r\boldsymbol{\xi}_m \tag{5}$$

where $\boldsymbol{A}^r$ and $\boldsymbol{b}^r$ are the $d \times d$ linear transformation matrix and the bias vector respectively associated to the regression class, r ($m \in r$). $\boldsymbol{\mu}_m$ and $\hat{\boldsymbol{\mu}}_m$ denote the original and adapted mean vectors respectively for component m. $\boldsymbol{W}^r = [\boldsymbol{A}^r \mid \boldsymbol{b}^r]$ and $\boldsymbol{\xi}_m = [\boldsymbol{\mu}'_m \; 1]'$ are the augmented transformation matrix and mean vector respectively. $\boldsymbol{W}^r$ can be estimated in an Expectation Maximization (EM) fashion by maximizing the following auxiliary function

$$\mathcal{Q}(\boldsymbol{W}^r) = K - \frac{1}{2}\sum_{m=1}^{M_r}\mathrm{Tr}(\boldsymbol{P}_m\boldsymbol{X}^{(mr)}) \tag{6}$$

where K subsumes terms independent of $\boldsymbol{W}^r$, M_r is the number of component in regression class r, $\boldsymbol{P}_m$ is a generic form of precision matrix for component m and $\boldsymbol{X}^{(mr)}$ is given by

$$\boldsymbol{X}^{(mr)} = \sum_{t=1}^{T}\gamma_m(t)(\boldsymbol{o}_t - \boldsymbol{W}^r\boldsymbol{\xi}_m)(\boldsymbol{o}_t - \boldsymbol{W}^r\boldsymbol{\xi}_m)'$$

$\gamma_m(t)$ is the posterior of component m at time t. For the case where $\boldsymbol{P}_m$ is full, direct optimization of equation (6) with respect to $\boldsymbol{W}^r$ using the direct closed-form solution [13] is computationally expensive and may result in numerical stability issue [13]. Instead, a simplified row-by-row estimation of $\boldsymbol{W}^r$ which guarantees an increase in the adaptation data likelihood may be used. This is achieved by differentiating equation (6) with respect to $\boldsymbol{w}_i^r$, the ith row of $\boldsymbol{W}^r$ and equating that to zero to yield the Maximum Likelihood (ML) solution as

$$\boldsymbol{w}_i^r = \boldsymbol{G}^{(rii)-1}\boldsymbol{k}^{(ri)}$$

where

$$\begin{aligned}
\boldsymbol{G}^{(rij)} &= \sum_{m=1}^{M_r} p_m(i,j)\boldsymbol{G}_m && (7)\\
\boldsymbol{k}^{(ri)} &= \sum_{m=1}^{M_r} \boldsymbol{p}_m(i)\boldsymbol{K}_m - \sum_{j=1, j\neq i}^{d} w_j^r\boldsymbol{G}^{(rij)} && (8)\\
\boldsymbol{G}_m &= \beta_m\boldsymbol{\xi}_m\boldsymbol{\xi}'_m && (9)\\
\boldsymbol{K}_m &= \boldsymbol{u}_m\boldsymbol{\xi}'_m && (10)
\end{aligned}$$

$p_m(i,j)$ and $\boldsymbol{p}_m(i)$ denotes the (i,j)th element and ith row of $\boldsymbol{P}_m$ respectively. The component level sufficient statistics are given by

$$\begin{aligned}
\beta_m &= \sum_{t=1}^{T}\gamma_m(t) && (11)\\
\boldsymbol{u}_m &= \sum_{t=1}^{T}\gamma_m(t)\boldsymbol{o}_t && (12)
\end{aligned}$$

This update formula is dependent on the other rows through the term $\boldsymbol{k}^{(ri)}$ in equation (8). Hence, an initial estimate of $\boldsymbol{W}^r$ is required and an iterative approach used. Although $\boldsymbol{W}^r$ can be initialized as an identity matrix, a better starting value may be found by using a diagonal precision matrix approximation, where $p_m(i,j) = 0$ for $j \neq i$. Equation (8) simplifies to that of a diagonal covariance matrix system [10].

$$\boldsymbol{k}^{(ri)} = \sum_{m=1}^{M_r} p_m(i,i)u_m(i)\boldsymbol{\xi}'_m \tag{13}$$

where $u_m(i)$ is the ith element of $\boldsymbol{u}_m$. In fact, the results presented later indicates that subsequent row-by-row iterations yield very little gain in terms of likelihood and the diagonal precision matrix approximation itself gives good estimates. This approximation approach is directly applicable to both SPAM and complexity control systems. For CAT system, on the other hand, MLLR mean adaptation is performed on the *effective* mean vector given by equation (2). Thus, the speaker-dependent CAT weights, $\lambda^{(s)}$, for each target speaker, s needs to be estimated before the standard MLLR mean adaptation approach is applied.

3.2. MLLR Covariance Adaptation

MLLR covariance adaptation is achieved via the following:

$$\hat{\boldsymbol{\Sigma}}_m = \boldsymbol{A}^r\boldsymbol{\Sigma}_m\boldsymbol{A}^{r'} \tag{14}$$

where $\boldsymbol{A}^r$ is the linear adaptation transformation matrix of the regression class r. $\boldsymbol{\Sigma}_m$ and $\hat{\boldsymbol{\Sigma}}_m$ are the original and adapted covariance matrices respectively. Equation (14) is equivalent to an STC model, which is a special case of the basis superposition precision matrix model as depicted in equation (3) where the bases are rank-1 matrices. There exists an efficient row-by-row update for the transformation matrix $\boldsymbol{A}^r$ as given by [14] where the ML solution is given by

$$\boldsymbol{a}_i^r = \boldsymbol{c}_i^r \boldsymbol{G}_i^{r-1} \sqrt{\frac{\beta}{\boldsymbol{c}_i^r \boldsymbol{G}_i^{r-1} \boldsymbol{c}_i^{r\prime}}} \tag{15}$$

where $\boldsymbol{a}_i^r$ is the ith row of $\boldsymbol{A}^r$, $\boldsymbol{c}_i^r$ is the vector of cofactors corresponding to $\boldsymbol{a}_i^r$ and

$$\boldsymbol{G}_i^r = \sum_{m=1}^{M_r} \sum_{t=1}^{T} \beta_m \lambda_{ii}^{(m)} (\boldsymbol{o}_t - \boldsymbol{\mu}_m)(\boldsymbol{o}_t - \boldsymbol{\mu}_m)' \tag{16}$$

$$\beta = \sum_{m=1}^{M_r} \sum_{t=1}^{T} \gamma_m(t) \tag{17}$$

From the above, it is obvious to see that a simple way to achieve covariance adaptation for basis superposition precision matrix models is to train speaker-dependent basis matrices. The efficiency of this kind of covariance adaptation depends on the computational cost of the basis matrix update of the precision matrix model. Unfortunately, this approach is computationally inefficient for SPAM and complexity control (multiple HLDA transforms) systems. Another way to achieve covariance adaptation is using the constrained MLLR adaptation where both the mean and covariance adaptation share the same linear transformation matrix. CMLLR will be described next.

3.3. Constrained MLLR Adaptation

Constrained MLLR (CMLLR) adaptation combines both MLLR mean and variance adaptation in a restrictive sense, such that the adaptation of the mean vector and the covariance matrix share the same linear transformation matrix. This constraint simplifies the adaptation to a feature-based speaker normalization scheme [15]. In CMLLR, a linear feature transformation matrix, $\boldsymbol{W}^r = [\boldsymbol{A}^r \mid \boldsymbol{b}^r]$, is estimated for each regression class, r such that

$$\hat{\boldsymbol{\zeta}}_t = \boldsymbol{A}^r \boldsymbol{o}_t + \boldsymbol{b}^r = \boldsymbol{W}^r \boldsymbol{\zeta}_t \tag{18}$$

where $\boldsymbol{\zeta}_t$ and $\hat{\boldsymbol{\zeta}}_t$ are the augmented vectors of the original and adapted observation respectively. The ML solution of $\boldsymbol{W}^r$ is found by maximizing the following auxiliary function

$$\mathcal{Q}(\boldsymbol{W}^r) = K + \beta \log |\boldsymbol{W}^r| - \frac{1}{2} \sum_{m=1}^{M_r} \mathrm{Tr}(\boldsymbol{P}_m \boldsymbol{X}^{(mr)}) \tag{19}$$

$\boldsymbol{W}^r$ is the transformation matrix, K subsumes terms independent of $\boldsymbol{W}^r$, and

$$\boldsymbol{X}^{(mr)} = \sum_{t=1}^{T} \gamma_m(t) (\boldsymbol{W}^r \boldsymbol{\zeta}_t - \boldsymbol{\mu}_m)(\boldsymbol{W}^r \boldsymbol{\zeta}_t - \boldsymbol{\mu}_m)'$$

Again, a row-by-row update approach is adopted here. Differentiating equation (19) with respect to $\boldsymbol{w}_i^r$, the ith row of $\boldsymbol{W}^r$, yields

$$\frac{\partial \mathcal{Q}(\boldsymbol{W}^r)}{\partial \boldsymbol{w}_i^r} = \beta \frac{\boldsymbol{c}_i}{\boldsymbol{c}_i \boldsymbol{w}_i^{r\prime}} - \boldsymbol{w}_i^{r\prime} \boldsymbol{G}^{(rii)} + \boldsymbol{k}^{(ri)} \tag{20}$$

where $\boldsymbol{c}_i$ is the cofactors of the ith row of $\boldsymbol{W}^r$. $\boldsymbol{G}^{(rij)}$ and $\boldsymbol{k}^{(ri)}$ are given by equations (7) and (8) respectively with the terms $\boldsymbol{G}_m$ and $\boldsymbol{K}_m$ given by

$$\boldsymbol{G}_m = \sum_{t=1}^{T} \gamma_m(t) \boldsymbol{\zeta}_t \boldsymbol{\zeta}_t' \tag{21}$$

$$\boldsymbol{K}_m = \boldsymbol{\mu}_m \boldsymbol{u}' \tag{22}$$

The sufficient statistics are $\beta, \boldsymbol{G}_m$ and $\boldsymbol{u}_m = \sum_{t=1}^{T} \gamma_m(t) \boldsymbol{\zeta}_t$. Setting equation (20) to zero yields the ML update for each row of $\boldsymbol{W}^r$ as

$$\boldsymbol{w}_i^r = \alpha \left(\boldsymbol{c}_i + \lambda \boldsymbol{k}^{(ri)} \right) \boldsymbol{G}^{(rii)-1} \tag{23}$$

Equation (23) is similar to the update formula derived for the case of diagonal covariance matrix [15], differed by the term $\boldsymbol{k}^{(ri)}$, which also depends on other rows in this case. α is found by solving a quadratic equation as described in [15]. It is easy to see that when $p_m(i,j) = 0$ for $j \neq i$, equation (23) simplifies to the case of diagonal covariance matrix systems.

CMLLR provides an alternative to variance adaptation. This is particularly useful for SPAM and complexity control systems where variance adaptation is computationally expensive. Unlike the case of MLLR mean, diagonal precision matrix approximation does not work for constrained MLLR because the estimated transforms operates on both the mean vectors and the precision matrices. However, the CMLLR transforms estimation process for SPAM models can be approximated using a diagonal covariance matrix model. For good approximation, this model should be the starting point used to train the SPAM model.

3.4. Sufficient Statistics for SPAM Models

The required statistics associated to each regression class r for both MLLR mean and CMLLR adaptations are $\boldsymbol{G}^{rij}$ for $1 \leq i \leq d$; $1 \leq j \leq i$ and $\boldsymbol{k}^{ri}$ for $1 \leq i \leq d$, as given by equations (7) and (8) respectively. The number of parameters to be stored for these statistics are $[\frac{d}{2}(d+1)]^2 + d^2$, which is dominated by $\boldsymbol{G}^{(rij)}$. For structured precision matrix models, the memory requirement can be reduced by exploiting the basis superposition structure. Substituting equation (3) into equation (7) yields

$$\boldsymbol{G}^{(rij)} = \sum_{b=1}^{n} s_b(i,j) \boldsymbol{G}^{(rb)} \tag{24}$$

$$\boldsymbol{G}^{(rb)} = \sum_{m=1}^{M_r} \lambda_{bb}^{(m)} \boldsymbol{G}_m \tag{25}$$

where $s_b(i,j)$ denotes the (i,j)th element of the bth basis matrix, $\boldsymbol{S}_b$ and $1 \leq b \leq n$. So, instead of storing $\frac{d}{2}(d+1)$ terms of $\boldsymbol{G}^{(rij)}$, only n terms of $\boldsymbol{G}^{(rb)}$ are needed. Thus, the required memory is reduced from the order $\mathcal{O}(d^4)$ to $\mathcal{O}(nd^2)$. These statistics are directly related to those presented in [1] where $\boldsymbol{G}_1^k$ and $\boldsymbol{G}_4^k$ are the same as $\boldsymbol{G}^{(rb)}$ for MLLR mean and CMLLR cases respectively. The notation k used in [1] has the same meaning as b used in this paper. Also, $\boldsymbol{G}_3^k$ relates to $\boldsymbol{K}^{(rb)} = \sum_{m=1}^{M_r} \lambda_{bb}^{(m)} \boldsymbol{K}_m$.

4. EXPERIMENTS AND RESULTS

4.1. Evaluation Framework

The evaluation framework used for system comparison was based on the the CU-HTK 10xRT evaluation system [16]. This is a multi-pass system uses sophisticated adaptation and CN based system combination. The overall system structure consists of two main stages: the initial lattice generation stage and the rescoring stage using multiple model sets. The confusion network outputs from different rescoring passes were finally combined. This is shown in figure 1. More details of the overall system architecture can be found in [16].

For both the CTS and BN systems the audio data was parameterized using 13 PLP features augmented with their first, second and third order derivatives. For the CTS systems only, Vocal Tract Length Normalization (VTLN) along with Cepstral mean and variance normalization was used in training and test. This 52 dimensional acoustic feature was projected down to 39 dimension using a global HLDA transform. Continuous density, mixture of Gaussians, cross-word triphone HMM systems were used for all systems and all acoustic models were built using discriminative training based on the minimum phone error (MPE) criterion [17]. Bandwidth-specific acoustic models were used for the BN task. Gender-specific BN models were also used for the non-adaptively trained system. However, all CTS acoustic models were gender independent.

The two baseline models used in the lattice rescoring (P3) stage were a SAT model employing constrained MLLR and an HMM set trained using a Single Pronunciation (SPron) dictionary. These model sets were adapted using lattice based MLLR in addition to standard adaptation based on the 1-best hypothesis.

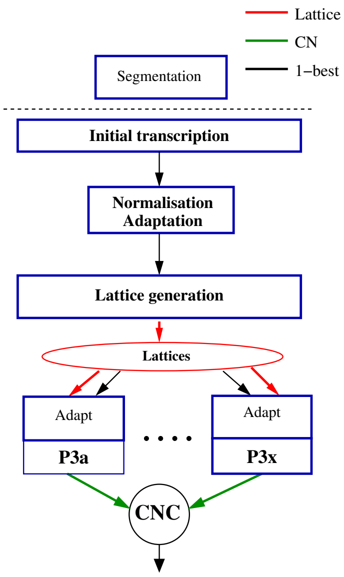

Fig. 1. CU-HTK 10xRT System

For both BN and CTS tasks a word-based 4-gram language model was trained on the acoustic transcriptions and additional Broadcast News data. The word-based 4-gram was then interpolated with a class-based trigram trained only on the associated acoustic transcriptions. The BN and CTS recognition dictionaries contain approximately 59k and 58k words respectively. Each word had about 1.1 pronunciations on average for both tasks.

4.2. Adaptation Results for SPAM models

This section presents the adaptation results for SPAM models based on the CTS and BN English tasks. Two forms of SPAM model were investigated in this work. The first was a standard SPAM system. The second was built within an adaptive training framework using CMLLR transforms. Instead of training the SAT+SPAM system from the SPAM system, the training approach described in [1] was adopted, where a speaker adaptively trained diagonal covariance matrix system (SAT+DIAGC) was used as the starting point. In other words, the SPAM precision matrix modelling was performed within the SAT feature space. In testing, MLLR mean transforms for the SPAM models were estimated using two row-by-row iterations as described in Section 3.1 (`mllr`) or simply approximated using the diagonal precision matrix assumption (`mllr+`). Similarly, the CMLLR transforms were estimated either using the exact method (`cmllr`) as described in Section 3.3 or approximated using a SAT+DIAGC system (`cmllr+`) for the adaptively trained SAT-SPAM system.

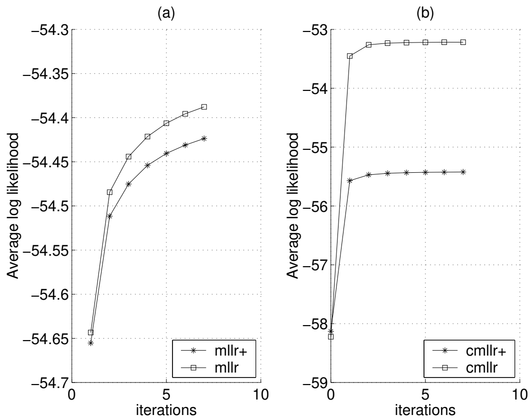

Fig. 2. Change in average log likelihood of one speaker on CTS with increasing number of MLLR iterations for (a) MLLR mean and (b) CMLLR, for 28-component SPAM model

Figure 2 illustrates the change in the average log likelihood of one speaker with increasing number of iterations for both MLLR mean and CMLLR adaptations. On each iteration, the component alignment was recomputed based on the transforms estimated in the previous iteration. The average log likelihood was found to increase upon every iteration. As expected the log-likelihood from the exact MLLR estimation (`mllr`) is always better than the approximate scheme (`mllr+`). However in both cases the likelihood increased at each iteration and the overall final difference was relatively small. For CMLLR, the log likelihood gain from using the `cmllr` method is about twice that of the approximated method, `cmllr+`, as depicted in Figure 2(b).

Word Error Rate (WER) performance was also examined. For the CTS task, 28-component models were trained using 400 hours of Fisher data (`fsh2004sub`) and evaluated on two test sets. `eval03` consists of two parts, Switchboard (`s25`) and Fisher (`fsh`), 3 hours each. `dev04`, on the other hand, is a 3 hours test set, containing only Fisher data. Table 1 summarizes the results of various adaptation configurations on CTS. The WERs of the baseline DIAGC system after MLLR adaptation were 22.3% and 18.4% on `eval03` and `dev04` respectively. SPAM model with diagonal

System	Adapt Config	eval03 s25	eval03 fsh	eval03 Avg	dev04 Avg
DIAGC	mllr	26.1	18.1	22.3	18.4
SPAM	mllr+	25.5	17.9	21.9	17.9
	mllr	25.5	18.0	21.9	18.0
SAT+DIAGC	cmllr	25.8	17.8	21.9	17.9
SAT+SPAM	cmllr+	25.0	17.6	21.4	17.6
	cmllr	24.9	17.5	21.3	17.5

Table 1. Comparisons of MLLR mean and CMLLR adaptations for 28-comp DIAGC and SPAM models on CTS system

precision matrix approximated MLLR adaptation gave 0.4-0.5% gains, although a large proportion of the gain on `eval03` came from `s25` (0.6%). Performing two additional row-by-row iterations, although improved the likelihood, degraded the WER performance by 0.1% on the `fsh` part of `eval03` and `dev04`. The SAT+DIAGC system is about 0.3%-0.5% absolute better than the non-SAT baseline on both test sets. Using this model to estimate the CMLLR transforms for the SAT+SPAM system (`cmllr+`) improved the WERs by 0.5% and 0.3% absolute on `eval03` and `dev04` respectively. Again, the gain on `s25` dominated for the `eval03` test set. Exact implementation using the `cmllr` method gave a consistent improvement of 0.1% on all test sets.

Next, a state-of-the-art SAT+SPAM system was trained using the 2180 hours `fsh2004h5etrain03b` training data. This training data comprises both Fisher (1820 hours `fsh2004`) and Switchboard (360 hours `h5etrain03b`) data. This system was evaluated on both `eval03` and `dev04` test sets and compared with the SAT+DIAGC system.

System	Adapt Config	eval03 s25	eval03 fsh	eval03 Avg	dev04 Avg
SAT+DIAGC	cmllr	22.7	15.5	19.2	16.1
SAT+SPAM	cmllr+	22.1	15.0	18.6	15.7
	cmllr	22.1	15.0	18.7	15.5

Table 2. Comparisons of CMLLR adapted 36-comp SAT+DIAGC and SAT+SPAM models on state-of-the-art CTS

In Table 2, the WER performance of the baseline SAT+DIAGC system was 19.2% and 16.1% on `eval03` and `dev04` respectively. As before, the difference between `cmllr` and `cmllr+` for SAT+SPAM is small. Comparing to SAT+DIAGC, the SAT+SPAM system gained about 0.5-0.6% and 0.4-0.6% absolute on `eval03` and `dev04` respectively. These gains were found to be statistically significant. Similar gains were also found with more complex adaptation techniques.

Similar comparisons were made on the BN task. 16-component models were trained using 374 hours of `bnetrain04sub` training data. This consists of 143 hours of carefully annotated data and 231 hours of lightly supervised data. Adaptation experiments were conducted based on three 3-hour test sets: `eval03`, `dev04` and `dev04f`. 4-gram rescoring lattices were generated using an adapted HLDA system[1]. Rescoring results are summarized in Table 3. For MLLR mean adaptation, a gender dependent (GD) DIAGC system was chosen as the baseline. This system gave WERs of 10.7%, 13.2% and 20.0% on the three test sets. The exceptionally poor performance on `dev04f` is due to the large mismatch between the training and the test data. Both `mllr+` and `mllr` configurations yielded the same performance, which is 0.1% absolute better than the baseline on `eval03` and `dev04`. The gain on `dev04f` is larger, 0.5% absolute. This shows that MLLR mean adaptation can be efficiently approximated with the diagonal precision matrix assumption for the SPAM models and other forms of precision matrix models such as EMLLT.

System	Adapt Config	Test Set WER (%) eval03	dev04	dev04f
DIAGC	mllr	10.7	13.2	20.0
SPAM	mllr+	10.6	13.1	19.5
	mllr	10.6	13.1	19.5
SAT+DIAGC	cmllr	10.6	13.1	19.5
SAT+SPAM	cmllr+	10.2	12.7	18.6
	cmllr	10.2	12.8	18.8

Table 3. Comparisons of MLLR mean and CMLLR adaptations for 16-comp DIAGC and SPAM models on BN system

Two forms of CMLLR adaptation for SAT+SPAM models were compared using the SAT+DIAGC system as the baseline. This system has the same WER performance as the MLLR mean adapted SPAM system. The `cmllr+` configurations gained 0.4% absolute on the first two test sets and 0.9% on `dev04f`. Again, there is a large gain from the adapted SPAM models due to the mismatch between the training and test sets. Similar performance was obtained on `eval03` using the exact `cmllr` configuration. Surprisingly, 0.1% and 0.2% degradations were observed on `dev04` and `dev04f` although the likelihood of the test data given these transforms was higher than those approximated using `cmllr+`. Apart from the gains from the `mllr+` and `mllr` SPAM models on `eval03` and `dev04`, all the gains shown in Table 3 were found to be statistically significant[2].

4.3. CTS Experiments

The CTS data set used for training, `fsh2004sub`, consists of 400 hours of Fisher conversations released by the LDC, with a balanced gender and line condition [18]. Quick transcriptions are provided by BBN, LDC and another commercial transcription service. Two CTS test sets were used for systems evaluation. A 6 hour DARPA RT-03 evaluation set, `eval03`, contains 72 conversations from the LDC Fisher collection, fsh, and Switchboard II phase 5, s25. Another DARPA development set `dev04` was also used, which includes 72 LDC released Fisher conversations. All CTS models have approximately 6k physical states after decision tree based tying. The number of components per state is 28 on average level.

Table 4 shows the baseline performance of the 10 time realtime CTS system. The 2-way combination between the SAT and SPron systems was the standard configuration used in the CUED CTS evaluation system. Significant error rate reduction over individual branches was achieved after system combination. The final error rates were 20.5% on `eval03` and 16.9% on `dev04`.

Table 5 shows the performances of various systems featuring techniques described in section 2. Note for the complexity control system the average number of components per state was 29.9

[1]Similar to the `P2` stage of the CU-HTK evaluation system

[2]Significance tests were carried out using the NIST Scoring Toolkit.

System		eval03			dev04
		s25	fsh	Avg	
P2-cn	HLDA	26.6	18.4	22.6	18.7
P3a-cn	SAT	24.5	17.1	20.9	17.3
P3c-cn	SPron	24.7	17.6	21.3	17.6
P3a+P3c		23.9	16.8	20.5	16.9

Table 4. CTS 10xRT system baseline performance

Gaussians per state and average number of retained dimensions per HLDA projection 42.6. The global HLDA system used for lattice generation was also re-adapted as a rescoring branch. The Gaussianization (GAUSS) and complexity controlled system (CTRL) systems gave marginal improvement. The SPAM system gave 0.8% absolute improvement on `eval03` over the HLDA baseline. An absolute word error reduction of 0.3% was also obtained on `dev04` against the P3b branch. Among all the adaptively trained systems, the SAT+SPAM outperformed all the other systems on both test sets. An absolute WER reduction of 0.4%∼0.5% were obtained on both sets over the SAT branch.

System		eval03			dev04
		s25	fsh	Avg	
P3b-cn	HLDA	24.8	17.7	21.4	17.5
P3d-cn	GAUSS	24.8	17.5	21.3	17.3
P3e-cn	CAT	24.9	17.2	21.2	17.5
P3g-cn	SPAM	24.1	16.9	20.6	17.2
P3h-cn	SAT+SPAM	23.9	16.9	20.5	16.8
P3i-cn	CTRL	24.5	17.5	21.1	17.6
P3d+P3c		24.1	17.0	20.5	17.0
P3e+P3c		24.2	16.8	20.7	17.0
P3g+P3c		23.6	16.5	20.2	16.8
P3h+P3c		23.6	16.4	20.1	16.6
P3i+P3c		23.9	16.8	20.5	16.9
P3c+P3d+P3h		23.6	16.4	20.1	16.5
P3c+P3h+P3i		23.3	16.3	19.9	16.6

Table 5. Extended CTS 10xRT system performance

The GAUSS, CAT, SPAM, SAT+SPAM and CTRL systems were then used for combination with the SAT and SPron systems. Using the SAT+SPAM branch reduced the error rate by 0.4% on `eval03` and 0.3% on `dev04`. The other approaches gave little gain over using the SAT system in combination with the SPron system. Adding the GAUSS system in a 3-way combination with the SPron and SAT+SPAM branches gave further marginal gain on `dev04`. Similarly the error rate on `eval03` was reduced by 0.2% using a 3-way combination including the CTRL system. To further increase the diversity and complimentary effects between different systems, a 6-way combination was performed. Unfortunately no further gain was obtained.

4.4. BN Experiments

The BN system was trained on 370 hours of training data. This consists of two parts [19], 140 hours of accurately transcribed broadcast news acoustic training data released by the LDC in 1996 and 1997 and 230 hours of data selected from the TDT4 audio corpora with close-captions based quick transcriptions. All BN models have approximately 7k physical states after decision tree based tying. The number of components per state is 16 on average level. Three BN test sets were used, each of them contains six 30 minutes broadcast news shows. The first set, `eval03`, was the DARPA RT-03 evaluation data set. It contains shows which were broadcast during February 2001. Two additional DARPA internal development sets, `dev04` and `dev04f` were also used. They contain shows of January 2001 and November 2003 respectively.

System		eval03	dev04	dev04f
P2-cn	HLDA	10.8	13.4	20.1
P3a-cn	SAT	10.3	12.9	18.7
P3c-cn	SPron	10.2	13.0	19.0
P2+P3a+P3c		10.1	12.6	18.6

Table 6. BN 10xRT system baseline performance

Table 6 shows the performance of the baseline BN 10xRT system. In contrast to the CTS system, a 3-way combination between the P2, P3a (SAT) and P3c (SPron) branches was the standard configuration used in CUED BN evaluation system. The final numbers for each of the tasks was 10.1%, 12.6% and 18.6%, with gains of 0.1% to 0.4% being obtained from system combination.

System		eval03	dev04	dev04f
P3b-cn	HLDA	10.5	13.1	19.5
P3d-cn	GAUSS	10.4	12.8	19.1
P3e-cn	CAT	10.4	12.8	19.1
P3g-cn	SPAM	10.2	12.7	18.8
P3h-cn	SAT+SPAM	10.1	12.5	18.5
P3i-cn	CTRL	10.5	12.8	19.3
P3e+P3c+P2		10.0	12.6	18.7
P3g+P3c+P2		10.0	12.6	18.5
P3h+P3c+P2		10.0	12.4	18.4
P3i+P3c+P2		10.1	12.6	18.8
P2+P3a+P3c+P3h		10.0	12.4	18.4

Table 7. Extended BN 10xRT system performance

Table 7 shows the performances of various BN systems. For the complexity control system there were an average of 16.5 components per state and 46.3 dimensions per HLDA transform. The Gaussianization system outperformed the HLDA system on all three sets. 0.3%∼0.4% error rate reduction is obtained on `dev04` and `dev04f`. The SPAM system was the best non-adaptively trained system, by an absolute WER reduction of 0.6%∼1.0% against the P3b system. Performances of the two SPAM systems are close. The CAT system consistently outperformed the gender-dependent HLDA baseline system on all sets. The gain from the CTRL system over the HLDA baseline was marginal similar to the CTS experiments. The SAT+SPAM system was then used for combination. Using the SAT+SPAM branch reduced the error rate by 0.1% on `eval03` and 0.2% on both `dev04` and `dev04f`, compared with the baseline 3-way combination configuration shown in table 6. Including the SAT system as additional branch in a 4-way combination with the P2, SPron and SAT+SPAM systems gave the same performance. Further marginal error rate reduction was obtained using a 7-way combination.

5. CONCLUSION

In this paper several advanced acoustic modelling techniques, Gaussianization, CAT, SPAM and complexity control were investigated for LVCSR training. Various MLLR-based adaptation schemes were discussed for these models, focusing primarily on the efficient row-by-row update approach for the MLLR mean and CMLLR adaptation of SPAM models. Performances of individual and combined systems were compared in the framework of a state-of-the-art 10 time real time system for both BN and CTS data. Experimental results show that these techniques are useful for further improving performance of current LVCSR systems.

6. REFERENCES

[1] S. Axelrod, V. Goel, B. Kingsbury, K. Visweswariah, and R. A. Gopinath, "Large vocabulary conversational speech recognition with a subspace constraint on inverse covariance matrices," in *Proc. Eurospeech*, 2003.

[2] G. Saon, A. Dharanipragada, and D. Povey, "Feature space gaussianization," in *Proc. ICASSP*, Montreal, Canada, 2004.

[3] S S Chen and R A Gopinath, "Gaussianization," in *Proceedings Neural Information Processing Systems*, 2000, pp. 423–429.

[4] M. J. F. Gales, "Cluster adaptive training of hidden markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 417–428, 2000.

[5] K. Yu and M. J. F. Gales, "Discriminative cluster adaptive training," Tech. Rep. CUED/F-INFENG/TR-486, Cambridge University Engineering Department, 2004.

[6] K C Sim and M J F Gales, "Precision matrix modelling for large vocabulary continuous speech recognition," Tech. Rep. CUED/F-INFENG/TR485, Cambridge University, 2004.

[7] G Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1973.

[8] A.R. Barron, J.J. Rissanen, and B Yu, "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, vol. 44, pp. 2743–2760, 1998.

[9] X. Liu and M. J. F. Gales, "Complexity control using marginalized discriminative growth functions," Tech. Rep. CUED/F-INFENG/TR-490, Cambridge University Engineering Department, 2004.

[10] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression speaker adaptation of contiuous density HMMs," *Computer Speech and Languages*, 1997.

[11] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Tech. Rep. CUED/F-INFENG/TR291, Cambridge University, 1997.

[12] K. C. Sim and M. J. F. Gales, "Basis superposition precision matrix modelling for large vocabulary continuous speech recognition," in *Proc. ICASSP*, 2004.

[13] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Languages*, vol. 10, pp. 249–264, 1996.

[14] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[15] M. J. F. Gales, "Maximum likelihood multiple projection schemes for hidden Markov models," Tech. Rep. CUED/F-INFENG/TR365, Cambridge University, 1999.

[16] G. Evermann and P. C. Woodland, "Design of fast LVCSR systems," in *Proc. ASRU*, St. Thomas, U.S. Virgin Islands, 2003.

[17] D. Povey and P. C. Woodland, "Minimum Phone Error and I-smoothing for improved discriminative training," in *Proc.ICASSP*, Orlando, Florida, 2002.

[18] G. Evermann, H. Y. Chan, M. J. F. Gales, B. Jia, D. Mrva, P. C. Woodland, and K. Yu, "Training LVCSR systems on thousands of hours of data," in *Submitted to ICASSP'05*, 2005.

[19] D. Y. Kim, M. J. F. Gales, G. Evermann, H. Y. Chan, K. C. Sim, D. Mrva, and P. C. Woodland, "Development of the CU-HTK 2004 broadcast news transcription system," in *Submitted to ICASSP'05*, 2005.