ISCA Archive
http://www.isca-speech.org/archive

ITRW on Adaptation Methods
for Speech Recognitionn
Sophia Antipolis, France
August 29-30, 2001

Speaker Adaptation for Continuous Density HMMs: A Review

P.C. Woodland

Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: pcw@eng.cam.ac.uk

Abstract

This paper reviews some popular speaker adaptation schemes that can be applied to continuous density hidden Markov models. These fall into three families based on MAP adaptation; linear transforms of model parameters such as maximum likelihood linear regression; and speaker clustering/speaker space methods such as eigenvoices. The strengths and weaknesses of each adaptation family are discussed along with extensions that have been proposed to improve the basic schemes which result in a number of hybrid approaches. A number of general extensions are discussed which include methods for improved unsupervised adaptation and discriminative adaptation. There is also a brief discussion of speaker normalisation and the relationship to model-based adaptation. The paper includes a brief discussion of other factors that directly interact with speaker adaptation of HMMs is included, such as adaptation to the acoustic environment and speaker-specific pronunciation dictionaries.

1. Introduction

Speaker adaptation has been an area of speech recognition technology that has attracted much attention over the last decade. While speaker independent (SI) speech recognition systems can show impressive performance, speaker trained or speaker dependent (SD) systems can provide an average word error rate (WER) a factor of two to three lower than an SI system if both systems use the same amount of training data. Hence the major rationale for investigating speaker adaptive (SA) systems is that they promise to produce a final system that has desirable SD-like properties but requires only a small fraction of the speaker-specific training data needed to build a full SD system.

SA systems can be aimed at improving the general performance level for all speakers, perhaps incrementally as more speech is available from a particular speaker. Adaptation can significantly improve the WER for outlier speakers such as non-natives or others not well represented in the SI training set. Furthermore SA techniques can be used to reduce the size (number of parameters) of the acoustic models required in a speech recognition system and the associated computational load compared to a non-adaptive SI system. This latter point can be very important for implementing a real system which will always have associated computational resource constraints.

Speaker adaptation systems operate in a number of *modes*. If the (word-level) transcription of the speaker-specific adaptation data is known then the adaptation is *supervised*, otherwise it is *unsupervised*: if the transcription is needed it must be estimated. While such an estimate may just be the errorful recognition output, some researchers have used confidence measures to ensure the adaptation process uses the most reliable material. Also adaptation modes are described as *static* (or block) in which all adaptation data is presented to the system before the final system is produced, or alternatively *dynamic* (or incremental) in which only part of the total adaptation data is available before use of the adapted system starts and the system continues to adapt over time.

The most appropriate mode of adaptation will depend on the application. For instance, enrollment of a new speaker in a speaker dependent dictation system will typically be a static supervised process, although this may then be followed by further supervised or unsupervised adaptation using the previously dictated text. Systems that are required to transcribe data in a non real-time manner often make use of multiple recognition passes and use unsupervised adaptation on the test data to improve the models for a subsequent recognition pass—this non-causal unsupervised style of adaptation is often termed *transcription* mode.

Speaker adaptation schemes should ideally be effective for small amounts of speaker-specific adaptation data and converge to a true speaker dependent estimate when a large amount of data is available. In many situations, if a large well-trained SI model is used, the baseline SI performance can be quite good and hence error-rate gains from speaker adaptation may be smaller than for rather simpler models. This is an important point to consider when evaluating speaker adaptation schemes where often the absolute performance (rather than relative gains) is the most important criteria!

The rest of this paper presents an overview of some of the most popular speaker adaptation schemes that are primarily applied to continuous density hidden Markov model (HMM) based speech recognition systems. Most of these are *model-based* and modify the parameters of the HMMs, however the effect of speaker normalisations schemes will also be mentioned. The model based adaptation schemes will be divided into three families: the maximum a posteriori (MAP) adaptation family; parameter transformation based adaptation using maximum likelihood linear regression (MLLR) and similar schemes; and a family related to speaker clustering methods or speaker-space methods. The strengths and weaknesses of these methods is considered and a number of hybrid schemes have appeared which combine the above methods in various ways. Extensions to the basic approaches are presented along with methods of improving unsupervised adaptation and discriminative adaptation methods. The relationship between model adaptation techniques for speaker-adaptation and environment adaptation are also discussed.

2. MAP family

Most HMM-based speech recognition systems are trained using maximum likelihood (ML) estimation: the parameter values, λ, are chosen so that the likelihood of the training data, $p(x|\lambda)$, is maximised. In maximum a posteriori parameter estimation (MAP) the parameters are set at the mode of the distribution $p(x|\lambda)p_0(\lambda)$ (the posterior distribution) where $p_0(\lambda)$ is the prior distribution of the parameters. The use of the prior distribution in MAP estimation means that less data is needed to get robust parameter estimates and hence it is a useful and widely used technique in speaker adaptation.

2.1. Standard MAP Approach

MAP estimation requires the definition of a prior distribution. It is convenient if the prior density is from the same family as the posterior distribution (the conjugate prior) if it exists. For mixture Gaussian HMMs such a conjugate prior of finite dimension does not exist and an alternative approach presented in [20] is usually used. For a particular Gaussian mean, with prior mean μ_0 the estimate is

$$\hat{\mu} \quad = \quad \frac{\tau\mu_0 + \sum_{t=1}^{T} \gamma(t) o_t}{\tau + \sum_{t=1}^{T} \gamma(t)}$$

where τ is a *meta-parameter* which gives the bias between the ML estimate of the mean from the data, and the prior mean, o_t is the adaptation vector at time t from a T length set and $\gamma(t)$ is the probability of this Gaussian at time t. Similar formulae [20] can be used to also update the mean and mixture weights in the system. Typically, values of τ between two and twenty are used.

One key advantage of the MAP approach is that as the amount of training data increases towards infinity the MAP estimate converges to the ML estimate. Its main drawback is that it is a *local* approach to updating the parameters i.e. only parameters that are observed in the adaptation data will be altered from the prior value. For large vocabulary applications, it not uncommon to use HMM systems with 10^5 Gaussians or more and in such circumstances the number of unobserved Gaussians (and unadapted by standard MAP) will be very large for small to moderate amounts of adaptation data and so adaptation with standard MAP can be very slow. This is a key problem which has led to other styles of adaptation to be investigated for large systems (and for very rapid adaptation for small systems also). It has also led to a number of extensions to MAP which aim to update unobserved parameters of the systems based on the observed data and hence increase the speed of adaptation.

2.2. Regression Based Model Prediction

The regression based model prediction (RMP) approach [1] (which is an extension of the approaches presented in [9] and [12]) aims to find correlations between the parameters of an HMM system and use these linear regression relationships to update poorly adapted or unobserved parameters based on well-adapted parameters.

First of all, a set of speaker dependent model sets are computed and for each Gaussian mean element in the system other mean values are found that are well correlated with its speaker-dependent changes. This search for correlated parameters is itself computationally demanding for a large system and requires constraints on the parameters searched for correlation. In use, RMP first updates the models using standard MAP, and then uses parameters that have received a reasonable amount of adaptation (source parameters) to generate parameter estimates for each unadapted or poorly adapted target value. The final mean value is a linear combination of the initial MAP estimate and the predicted value (weighted in accordance with the estimated inverse variances).

In practice, RMP converges to the same error rate as MAP, but out-performs MAP for small amounts of adaptation data. For instance for just one three second adaptation sentence on the 1000 word Resource Management task using a 6 Gaussian per state, state clustered triphone system a 8% reduction in WER, while MAP gives no improvement since only about 5% of the parameters in the system receive any adaptation data [1].

2.3. Structural MAP

A rather different type of augmented MAP approach to tackle the same speed of adaptation problem is termed structural MAP (SMAP) [35]. The Gaussians in the system are all organised into a tree structure and then a mean offset and a diagonal variance scaling term are recursively computed for each layer of the tree starting at the root node (containing all the Gaussians) and then descending the tree. At each level in the tree, the distribution from the node above is used as a prior. It is shown in [35] that the use of this approach increases the speed of adaptation over standard MAP while converging to the MAP solution as the amount of adaptation data is increased.

3. Linear Transformation Family

An alternative approach to the speaker adaptation problem is to estimate a linear transformation of the model parameters (or sometimes the observation vectors) to construct a more appropriate model. The advantage of this approach is that the same transformation can be used for a large number of (or even all) Gaussians in an HMM system and this sharing of transformation parameters provides a route towards (fairly) rapid adaptation. Furthermore if relatively few parameters are to be estimated then the process will be robust and unsupervised adaptation can be used.

There are a number of schemes that use linear transformations. First the popular maximum likelihood linear regression (MLLR) scheme will be described which is an unconstrained transformation, followed by constrained transformations, efficient full variance transforms and speaker adaptive training based on MLLR.

3.1. Maximum Likelihood Linear Regression

In basic MLLR [28] the Gaussian mean parameters are updated according to

$$\hat{\mu} = \mathbf{A}\mu + b$$

where $\mathbf{A}$ is an $n \times n$ matrix and b is an n dimensional vector (and n is dimensionality of the observations). This equation is sometimes written as

$$\hat{\mu} = \mathbf{W}\xi$$

where $\mathbf{W}$ is an $n \times (n+1)$ matrix and ξ is the extended mean vector

$$\xi^T = \left[\begin{array}{cccc} 1 & \mu_1 & \dots & \mu_n \end{array} \right]$$

In MLLR, the transformation matrix $\mathbf{W}$ is estimated such that the likelihood of the adaptation data is maximised. It can be shown [28] that there is a closed form solution to the $\mathbf{W}$ matrix estimation problem using, as usual, the Expectation-Maximisation (E-M) algorithm [10]. Furthermore, in many circumstances (where the initial models can provide good Gaussian-frame alignments) only a single iteration of E-M is required to estimate the matrix. Usually there are many Gaussians per matrix i.e. the transformation matrix is tied over a number of Gaussians. This transform sharing can allow all the Gaussians in a system to be updated with only a relatively small amount of adaptation data.

However, there is a tradeoff between robust adaptation via a global transform and using precise transforms that apply to a smaller number of e.g. phone-specific Gaussians. One solution that allows a good compromise to be drawn is to use a *Regression Class Tree* [27]. The idea is to arrange so that Gaussians that are close in acoustic space are clustered together and always undergo the same transformation (these groups are known as base classes). If the clustered components are then arranged into a tree structure (with all at the root node), then, depending on the amount of adaptation data available the tree may be descended to an appropriate depth and a set of transformations generated where each transformation will be for a set of base classes.

The MLLR mean transformation was presented for the case when the $\mathbf{A}$ matrix is full. However for feature vectors consisting of static parameters, 1st and 2nd differentials, approximately the same performance per transform can be obtained using block-diagonal matrices [29]. For still smaller amounts of adaptation data diagonal matrices can be used, although for the same number of transforms these are far less effective than full or block-diagonal matrices. However this is done precisely, it should be noted that thresholds need to be set to ensure robust transform estimation. A typical threshold for the estimation of a full-matrix for 39 dimensional observations would be 1000-1500 frames. Smaller threshold values can be used in cases where the transformation is block-structured or even diagonal. and these techniques allow the sensible use of MLLR adaptation (with simpler, less powerful transform structures) even for rather small amounts of adaptation data.

MLLR transforms can also be estimated in incremental adaptation mode [27]. This simply requires that the sufficient statistics needed to compute the transforms continue to be updated and then the mean transform can be recalculated at any time. In the case that the adaptation process has not altered the Gaussian-frame occupation probabilities then the same transformed system will result as using static adaptation.

While the most important speaker specific effect concerns the Gaussian means, the Gaussian variances can also be updated [13, 15]. The variance transforms $\mathbf{H}$ are found after the mean transforms have been estimated. Originally the form

$$\hat{\Sigma} = LHL^T$$

was used where L is the Choleski factor of the original covariance matrix Σ. For the case of a diagonal variance transform (with a simple bias for the mean) this is the same as the variance transform suggested in [29].

A variance transform of the form

$$\hat{\Sigma} = H\Sigma H^T$$

is proposed in [15] which has the advantage that it can be applied efficiently by transforming the mean parameters and the observations even for full variance transformations. However the transformation elements need to be estimated using an iterative procedure given the sufficient statistics.

Typically mean-only MLLR gives a 15% reduction in WER on large vocabulary clean speech tasks over the most accurate speaker independent models available using about a minute of adaptation data, and speaker dependent performance can often be achieved with perhaps thirty minutes of speech and many adaptation transforms [28]. If the technique is used to adapt a system with severe speaker mismatch, large reductions in WER can result. In this latter case there is some benefit from running multiple E-M iterations to update the adaptation matrices to refine the Gaussian-frame alignments [27].

3.2. Constrained MLLR

The MLLR formulation described above estimates independent transforms for the means and the variances. The constrained transform case, which was introduced in [11] for the diagonal transform case and extended in [15] to full transforms is of the form

$$\begin{aligned}\hat{\mu} &= \mathbf{A_c}\mu - b_c \\ \hat{\Sigma} &= \mathbf{A_c}^T\Sigma\mathbf{A_c}\end{aligned}$$

This can be convenient since this is equivalent to transforming the observation vectors such that the vector at time t becomes

$$\hat{o_t} = \mathbf{A_c}^{-1}o_t + \mathbf{A_c}^{-1}b_c$$

noting that a factor of $|\mathbf{A_c}|$ is also needed when calculating the Gaussian likelihood. The maximum likelihood solution for this form requires iterative optimisation given the sufficient statistics, but gives similar performance to using standard unconstrained MLLR with the same form of transformation matrix.

3.3. Speaker Adaptive Training

For the transformation based schemes presented above it has been assumed that the original "seed" models to be transformed are speaker independent. However, somewhat improved performance can be obtained by constructing models specifically aimed at speaker adaptation. This leads to the speaker adaptive training (SAT) approach [2] which estimates the parameters of the seed models by training MLLR mean transforms for each of the training set speakers and then estimating the mean, variance and mixture weights of the models with these speaker specific transforms in place. The estimation of the adaptation transforms and the seed (or canonical) model parameters would normally be done in an interleaved fashion. SAT tends to result in models with significantly higher training set likelihoods and smaller estimated variances and can result in noticeably improved recognition results with adaptation [2, 33]. However the statistics required for this form of SAT require a considerable amount of storage and increased computation.

An alternative, and easier to implement variant of SAT is to use a constrained MLLR transform (rather than standard unconstrained MLLR) and transform the incoming data, rather than the model means. In this case, the canonical model estimation formula is almost unchanged [15]. An additional advantage of using constrained MLLR SAT training is that it allows SAT to be applied when also using a discriminative training method to estimate the canonical model parameters.

3.4. Improving MLLR Robustness

As mentioned above, it is necessary to have sufficient data points to robustly estimate MLLR or similar transforms. If appropriate thresholds/forms of transforms are not used then poor performance (even poorer than speaker independent performance) can be the result due to over-training on the adaptation data. Of course, in practice, limits on the form of the transforms combined with a range of thresholds are used to ensure that performance is never poorer than SI. However it would be preferable to instead use schemes which don't, for instance, require switching from diagonal transforms to block-diagonal as more data becomes available.

Various solutions to this problem have been suggested and all increase the applicability for MLLR for rapid adaptation. The solutions include a somewhat ad-hoc MAP-like interpolation between the original mean and the MLLR estimated mean [21]. In both [7] and [8], it is suggested that a prior distribution for the mean transformation matrix parameters be used (dubbing the technique MAPLR) and this improves performance when very small amounts of data are available. The prior distribution can be estimated by generating transforms from the set of training speakers. The MAPLR method can be further extended by using an SMAP prior distribution which leads to SMAPLR [36].

The above approaches all use a MAP-style estimation approach for MLLR parameters. Alternatively a variant of the E-M algorithm that optimises a discounted likelihood criterion and doesn't quickly over train was suggested in [22]. This DLLR technique also improves robustness for small amounts of adaptation data when many transforms are to be trained.

4. Speaker Clustering/Speaker Space Family

The previous approaches have not explicitly used information about the characteristics of an HMM set for particular speakers (although some of this information is used in techniques like RMP at a parameter level).

The simplest instance of such an approach is the use of gender dependent models which are widely used in speaker independent systems. Traditional speaker clustering (e.g. [34]) goes a step further and estimates HMMs for a number of speaker groups. However the problem with this type of approach is that by taking hard decisions about speaker type, the training data is fragmented and it is possible to make a poor choice of speaker group when in use.

Recently there has been interest in the cluster adaptive training (CAT) [16] and eigenvoice techniques [25] which can be viewed as generalisations of this idea. These both form a weighted sum of “speaker cluster” HMMs, and use this interpolated model to represent the current speaker. The parameters of the sets of cluster models that are estimated can be viewed as representing the axes of a “speaker space” and then the mean vectors[1] for a particular speaker are found by estimating the appropriate point for the speaker in this speaker space. One of the major differences between the CAT and eigenvoice approaches is how the cluster models are estimated.

4.1. Cluster Adaptive Training

The aim of CAT [16] is to represent a speaker as a weighted sum of individual speaker cluster models. It is assumed that all the different speaker cluster models have a common variance and mixture weights and only the Gaussian mean values vary. Thus the means for a particular Gaussian for a particular speaker is found as

$$\hat{\mu} = \sum_c \lambda_c \mu_c$$

where the parameters of the model are the speaker-specific λ_c which define the cluster weights (or the points in “speaker-space”) and μ_c is the the corresponding mean of corresponding Gaussian in cluster c i.e. the canonical model for cluster c. Note that cluster models each have the same number of parameters as the target model (and often as a speaker independent model). However the approach gains its data efficiency from the fact that only very few parameters (the λ_c) need to be estimated for each speaker. For a particular set of canonical speaker cluster models, and some adaptation data, maximum likelihood weight estimation formulae for the cluster weights can be derived. Furthermore given sets of weights for individual speakers the canonical speaker cluster model means can also be updated. Therefore the CAT canonical model estimation scheme consists of interleaving weight estimation and speaker cluster updates for the training data. Given a transcription of the adaptation data, the same weight estimation approach is used for test-speakers.

The above scheme describes “model-based” CAT in which the cluster models are estimated directly. An alternative form, also discussed in [16] is to use MLLR transforms from a “canonical model” to represent each of the individual speaker clusters, and an advantage is that the total number of parameters set in is reduced relative to model-based CAT with the same number of clusters. Transform-based CAT estimation therefore combines SAT training of MLLR transforms with the CAT estimation process.

Experiments in [16] on a large vocabulary dictation task show that CAT modelling can reduce the WER by 7% using a single adaptation sentence and two speaker clusters. In this case one of the clusters was the SI model which is referred to as a “bias” cluster of fixed weight. The reduction in WER increases slightly if more adaptation sentences and speaker clusters are used. When 8 clusters were used with the most complex HMMs, transform-based and model-based CAT gave similar performance.

When there are different types of speaker (e.g. accent groups; male/female; or speakers present in different noise conditions) the model-based CAT cluster models (or the transforms in transform-based CAT) can be initialised using models trained from reduced data sets. One alternative when more clusters are required is to use speaker clustering. Another possibility suggested in [16] is to use an initialisation based on eigenvoices (see discussion in Section 4.2).

4.2. Eigenvoices

The eigenvoice technique [25] also performs speaker adaptation by forming models as a weighted sum of canonical speaker HMMs and adapts just the mean vectors. However, the eigenvoice method finds these canonical speakers (eigenvoices) using principal component analysis (PCA) of sets of “supervectors” constructed from all the mean values in a set of speaker dependent HMM systems. The eigenvoices with the largest eigenvalues are chosen as a basis set. During adaptation the maximum likelihood eigen-decomposition algorithm is proposed in [25] to estimate the weighted combination of eigenvoices. Of course this algorithm is identical to the CAT weight estimation algorithm for adaptation. Furthermore, the same weight estimation formulae can be derived on the basis of using a weighted projection technique [41].

In [25], the eigenvoice technique was evaluated for a small vocabulary task using simple HMM models and produces impressive performance with small amounts of data. Unfortunately for large HMM systems (with perhaps 100,000 Gaussians) the construction of separate HMM systems for all speakers and subsequent PCA analysis is particularly difficult. There are two main issues here: firstly if mixture distributions are used then these must be “aligned” between the various sets of models and secondly the number of parameters to estimate in the full speaker dependent models. This can result in both estimation issues and storage problems. Several solutions have been proposed for these problems.

[1] The variance and mixture weight parameters are not normally re-estimated in these approaches

To overcome the problem of estimating many SD models, a direct maximum likelihood solution to finding the canonical speakers is proposed in [30]. Not surprisingly this is very similar to the CAT estimation procedure for the cluster models. In [4], MAP and MLLR is used to create the speaker dependent models of complete SD model sets for each speaker before PCA. An alternative which was used in the CAT eigenvoice initialisation, [16], is to build a small HMM set for each speaker and perform PCA analysis on the SD means. Then the weight vector from PCA for the retained eigenvoices each training speaker along with the *a posteriori* probability of Gaussian occupation for each frame of data from an SI system allows direct estimation of the complex HMMs for each speaker cluster. This is an elegant solution avoiding the problems of estimating many large SD HMM sets.

While CAT and other speaker-clustering/speaker space schemes can give good performance for small amount of adaptation data, one issue is that performance does not continue to improve as more data is made available. One solution to this problem is to use a prior distribution for MAP based on eigenvoices [5]. This allows both rapid adaptation and long-term convergence to the MAP solution.

5. Further Extensions

This section briefly describes a number of extensions of the basic techniques described above. Also included here are methods to improved unsupervised adaptation and discriminative adaptation.

5.1. Extensions for Transform-Based Adaptation

Over the last few years, several papers have investigated what might be generically called "multiple-cluster" adaptation schemes. This has the motivation that the transform for a particular set Gaussian might be estimated using a linear combination of transforms estimated for broader sets Gaussians. The general idea is to increase the number of effective transforms in the system while only increasing the number of parameters to be estimated by a small amount (i.e. the extra transform interpolation weights). [14]. An related approach, maximum likelihood stochastic transforms [6], interpolates Gaussian likelihoods from sets of transforms. There are many other forms of multiple cluster schemes that can use either linear transform adaptation or CAT-style models. An overview of some of the possibilities for multiple cluster adaptation are given in [17].

Another extension of the linear transform approach that has achieved some attention is the use of non-linear transformations of the model parameters. However work reported to date has only shown rather minor improvements over linear transform methods.

5.2. Improving Unsupervised Adaptation

Normally when speaker adaptation schemes are used in an unsupervised adaptation mode, the adaptation supervision word sequence is computed using the output of a speech recogniser. The adaptation data is either previously acquired data for the speaker or in the case of transcription mode adaptation in a multi-pass system, the block of data currently being processed. Many adaptation schemes simply apply the adaptation supervision as if it were correct although, in general, it will contain recognition errors. In this case the issue of robust of adaptation parameter estimation is crucial and techniques such as MLLR and speaker-space methods are preferred.

One approach to improve unsupervised adaptation performance is to use word correctness confidence scores and only use data for adaptation which has a high enough probability of correctness. This type of scheme has been studied in [3, 38, 40, 44]. While in some circumstances improvements in adaptation performance can be shown, the technique reduces the amount of data available for adaptation.

An alternative to direct use of confidence scores is to perform adaptation using a set of alternative hypotheses. One implementation of this idea was developed for MLLR adaptation in [31] with a closely-related technique described in [38]. These collect the posterior probabilities of Gaussian occupation for MLLR from a forward pass through the lattice of recognition alternatives. This technique automatically weights different alternatives and hence doesn't discard complete frames of data. It is shown in [38] that this can lead to the beneficial estimation of a larger number of MLLR transforms. While this lattice-based approach has so-far been applied to MLLR adaptation, it could also be applied for many of the other adaptation techniques discussed in this paper.

5.3. Discriminative Adaptation

The techniques described in this paper estimate parameters to increase either the likelihood of the adaptation data (MLLR, CAT) or to find the MAP estimate of the parameter values. Neither of these approaches is directly aimed at reducing the word error rate of the adaptation data. Recently there has been renewed interest in discriminative training of the SI parameters of large vocabulary speech recognition systems [43]. There is also a recent trend towards performing discriminative adaptation.

In [19] a discriminative version of MAP was used while in [40] the frame discrimination criterion was used to estimate linear transform parameters. In [39] discriminative linear transforms were estimated using an objective function that is an interpolation of maximum likelihood and maximum mutual information (MMI) for linear transform estimation. It was shown that this can provide significant reductions in WER for supervised adaptation of a system trained on native speakers of English to non-natives. A method equivalent to using the MMI objective function for transform parameters was presented in [23] for unsupervised adaptation.

6. Speaker Normalisation

All of the above techniques have modified the models in some way (although a single transform for constrained MLLR can be implemented just by modifying the observation features). Furthermore all of the above techniques require a word level transcription (either known or estimated by a recogniser to operate). This in itself imposes certain constraints on the adaptation system.

Speaker normalisation techniques solely alter the observation features to the system. The simplest of these is the widely used cepstral mean normalisation which subtracts the long term cepstral mean from individual speakers and has some (small) speaker normalisation effect. An alternative speaker normalisation technique that has become increasingly popular is vocal tract length normalisation (VTLN). VTLN re-scales to the frequency axis with the aim of accounting for the difference in vocal tract length between speakers. A grid search over possible frequency warpings is often used to select the factor which maximises the likelihood of the data [26]. For the purpose of finding the warp factor a Gaussian mixture model is often used which doesn't require a word-level transcription. On clean speech tasks a reduction in WER due to VTLN of 10% is typical.

It has been reported by several authors (e.g. [33]) that the beneficial effects from VTLN and MLLR are largely additive. This apparently surprising result was further investigated in [37] in which the VTLN process is approximated using linear cepstral transforms. It was shown there that while this additive property holds for unconstrained MLLR, if constrained MLLR is used, there is no additional gain from using both VTLN and MLLR.

7. Relationship to Environmental Adaptation

The adaptation techniques described above all modify the acoustic models to better match some adaptation data. Certainly, for the general MAP and MLLR type techniques there is no in-built speaker model and the techniques can also be used effectively for cases of acoustic environment mismatch (channel or noise) as well as the frequent case of combined speaker and channel adaptation.

In cases of severe channel mismatch when operating in unsupervised adaptation mode it has been found to be beneficial to employ multiple iterations of decoding and adaptation [42]. Furthermore in systems which use multiple passes and generate word-lattices, smaller and more accurate lattices can be obtained if adaptation is performed before lattice generation [42].

One interesting issue is distinguishing between speaker mismatch and environment mismatch. This would be useful to provide a speaker adapted system that was independent of the acoustic environment. In theory, VTLN is able to just compensate for speaker mismatch, although most of the warp factor estimation algorithms are affected by channel effects also. In [30] it is suggested that the eigenvoice method can be used to model speaker changes and then use MLLR adaptation to account for environment mismatch.

Another approach to combined speaker/environment adaptation is discussed in [18] in which different noise environments are modelled by CAT/eigenvoices and speaker adaptation is performed by MLLR. The work reported in [18] tries to factor out these different sources of variability in an adaptive training framework. Methods of then combining speaker and acoustic environment models to form an HMM set for a particular speaker in a particular acoustic environment are discussed. This approach may lead to rapid combined speaker and environment adaptation and allow speaker transform parameters to be retained across noise conditions.

8. Summary and Outlook

An overview of some of the most widely used techniques in speaker adaptation have been described. These can be split into several families and the major properties of each type have been given.

The main advantage of the MAP method is its sound theoretical basis and convergence to speaker dependent performance with increasing data, however the major drawback is the speed of adaptation for large HMM systems. This latter point can be improved somewhat by using techniques such as RMP and SMAP.

The transformation based approaches work well after a few sentences of adaptation data have been made available and can work reliably in all adaptation modes. The approach does not need any prior information about the distribution of speaker (or environment) types to be effective. One disadvantage of this type of technique is that since there is no model of speaker variation any acoustic mismatch will be modelled. A variety of approaches have been suggested to prevent over-training.

The speaker clustering family looks for either a choice of, or an interpolation between, canonical speaker models. The CAT/eignvoice approaches can be viewed as finding appropriate points in a speaker space. This type of approach can be effective for a small amount of adaptation data but the gains available with added data are more limited and in such cases a technique such as MLLR or MAP may be preferable.

All of the techniques described above have some particular strong points and in some cases work has been done to explicitly combine them, for instance using initial MLLR adaptation as a prior for later MAP adaptation as more data is available. It has been a recent trend to try and combine the strengths from different methods and this can yield methods that scale well over a wide range of adaptation data.

Speaker normalisation techniques are in some ways complementary to model-based adaptation and indeed near additive gains have been reported from using unconstrained MLLR in combination with VTLN.

The model based adaptation techniques can also be applied to environment adaptation, but only a small amount of work has been done to explicitly separate environmental adaptation effects from speaker effects. This remains a very interesting area for future work.

While all of the above techniques directly effect the HMM parameters, in situations where there is severe pronunciation mismatch, adaptation of the pronunciation dictionary may also be beneficial. For instance, [24] that hen using a system mismatched for accent, reductions in WER due to pronunciation dictionary adaptation alone of 19% were shown and when combined with MLLR a total reduction of 40% in WER was obtained. Therefore it is interesting that as for environmental adaptation, pronunciation adaptation can be usefully combined with and acoustic model adaptation.

In future, it will be necessary to improve speaker adaptation systems by incorporating more extensive knowledge of speaker variation at both the acoustic and the pronunciation level. After all, while the progress in speaker adaptation over the last decade has been impressive, we still have a long way to go to match the seamless integration of extremely rapid, unsupervised incremental adaptation in the human perception system.

9. References

[1] S.M Ahadi & P.C. Woodland (1997) Combined Bayesian and Predictive Techniques for Rapid Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech & Language*, Vol. 11, pp. 187-206.

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

[3] T. Anastasakos & S.V. Balakrishnan (1998). The Use of Confidence Measures in Unsupervised Adaptation of Speech Recognizers. *Proc. ICSLP'98*, pp. 2303-2306, Sydney.

[4] H. Botterweck (2000). Very Fast Adaptation for Large Vocabulary Speech Recognition Using Eigenvoices. *Proc. ICSLP'2000*, Vol IV, pp. 354-357, Bejing.

[5] H. Botterweck (2001). Anisotropic MAP Defined by Eigenvoices for Large Vocabulary Speech Recognition. *Proc. ICASSP'2001*, Salt Lake City.

[6] C. Boulis, V. Diakoloukas & V. Digalakis (2001). Maximum Likelihood Stochastic Transformation Adaptation for Medium and Small Data Sets. *Computer Speech & Language*, Vol. 15, pp. 257-285.

[7] C. Chesta, O. Siohan & C.H. Lee (1999). Maximum A Posteriori Linear Regression for Hidden Markov Model Adaptation. *Proc. Eurospeech'99*, pp. 211-214, Budapest.

[8] W. Chou (1999). Maximum A Posteriori Linear Regression with Elliptically Symmetric Matrix Priors. *Proc. Eurospeech'99*, pp. 1-4, Budapest.

[9] S.J. Cox (1995). Predictive Speaker Adaptation in Speech Recognition. *Computer Speech & Language*, Vol. 9, pp. 1-17.

[10] A.P. Dempster, N.M. Laird & D.B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Vol. 39, pp. 1-38.

[11] V. Digilakis, D. Ritchev & L. Neumeyer (1995) Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures. *IEEE Trans. SAP*, Vol. 3, pp. 357-366.

[12] S. Fururi (1980). A Training Procedure for Isolated Word Recognition Systems. *IEEE Trans. ASSP*, Vol 28, pp. 129-136.

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

[14] M.J.F. Gales (1997). Transformation Smoothing for Speaker and Environmental Adaptation. *Proc. Eurospeech'97*, pp. 2067-2070, Rhodes.

[15] M.J.F. Gales (1998). Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. *Computer Speech & Language*, Vol. 12, pp. 75-98.

[16] M.J.F. Gales (2000) Cluster Adaptive Training of Hidden Markov Models. *IEEE Trans. SAP*, Vol. 8, No. 4, pp. 417-428.

[17] M.J.F. Gales (2001) Multiple Cluster Adaptive Training Schemes. *Proc. ICASSP'2001*, Salt Lake City.

[18] M.J.F. Gales (2001) Acoustic Factorisation: Theory and Initial Evaluation. *Technical Report, CUED/F-INFENG/TR.419*, Cambridge University Engineering Dept.

[19] Y. Gao, B. Ramabhadran & M. Picheny (2000). New Adaptation Techniques for Large Vocabulary Continuous Speech Recognition. *Proc. ISCA ITRW ASR2000*, Paris.

[20] J.L. Gauvain & C.H. Lee (1994) Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Trans. SAP*, Vol. 2, pp. 291-298.

[21] S. Goronzy & R. Kompe (1999) A MAP-Like Weighting Scheme for MLLR Speaker Adaptation. *Proc. Eurospeech'99*, pp. 5-8, Budapest.

[22] A. Gunawardana & W. Byrne (2001). Discounted Likelihood Linear Regression for Rapid Speaker Adaptation. *Computer Speech & Language*, Vol. 15, pp. 1-14.

[23] A. Gunawardana & W.J. Byrne (2001). Discriminative Adaptation with Conditional Maximum Likelihood Linear Regression. *Presented at NIST Hub5 Workshop*, May 2001.

[24] J.J. Humphries & P.C. Woodland (2001) Accent Modelling and Adaptation in Automatic Speech Recognition. To appear, *IEEE Trans. SAP*.

[25] R. Kuhn, J.C. Junqua, P. Nguyen, & N. Niedzielski (2000) Rapid Speaker Adaptation in Eigenvoice Space. *IEEE Trans. SAP*, Vol. 8, No. 6, pp. 695-707.

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

[27] C.J. Leggetter & P.C. Woodland (1995). Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression. *Proc. ARPA Spoken Language Technology Workshop*, pp. 104-109. Morgan Kaufmann.

[28] C.J. Leggetter & P.C. Woodland (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech & Language*, Vol. 9, pp. 171-185.

[29] L. Neumeyer, A. Sankar & V. Digilakis (1995). A Comparative Study of Speaker Adaptation Techniques. *Proc. Eurospeech'95*, pp. 1127-1130, Madrid.

[30] P. Nguyen, C. Wellekens & J.C. Junqua (1999). Maximum Likelihood Eigenspace and MLLR for Speech Recognition in Noisy Environments. *Proc. Eurospeech'99*, pp. 2519-2522, Budapest.

[31] M. Padmanabhan, G. Saon & G. Zweig (2000). Lattice-Based Unsupervised MLLR for Speaker Adaptation. *Proc. ISCA ITRW ASR2000*, pp. 128-131, Paris.

[32] M. Pitz, F. Wessel & H. Ney (2000) Improved MLLR Speaker Adaptation Using Confidence Measures for Conversational Speech Recognition. *Proc. ICSLP'2000*, Bejing.

[33] D. Pye & P.C. Woodland (1997) Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition. *Proc. ICASSP'97*, pp. 1047-1050, Munich.

[34] . T. Kosaka & S. Sagayama (1994). Tree-structured Speaker Clustering for Fast Speaker Adaptation. *Proc. ICASSP'94*, pp. 245-248. Adelaide.

[35] K. Shinoda & C.H. Lee (1997). Structural MAP Speaker Adaptation Using Hierarchical Priors. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 381-388, Santa Barbara.

[36] O. Siohan, T.A. Myrvoll & C.H. Lee (2000) Structural Maximum A Posteriori Regression for Fast HMM Adaptation. *Proc. ISCA ITRW ASR2000*, pp. 120-127, Paris.

[37] L.F. Uebel & P.C. Woodland (1999). An Investigation into Vocal Tract Length Normalisation. *Proc. Eurospeech'99*, pp. 2519-2522, Budapest.

[38] L.F. Uebel & P.C. Woodland (2001). Speaker Adaptation Using Lattice-based MLLR. *Proc. ISCA ITR-Workshop on Adaptation Methods in Speech Recognition*, Sophia-Antipolis.

[39] L.F. Uebel & P.C. Woodland (2001). Discriminative Linear Transforms for Speaker Adaptation. *Proc. ISCA ITR-Workshop on Adaptation Methods in Speech Recognition*, Sophia-Antipolis.

[40] F. Wallhoff, D. Willett & G. Rigoll (2000). Frame-Discriminative and Confidence-Driven Adaptation for LVCSR. *Proc. ICASSP 2000*, pp. 1835-1838, Istanbul.

[41] R.J. Westwood (1999) *Speaker Adaptation Using Eigenvoices.* MPhil Thesis, University of Cambridge.

[42] P.C. Woodland, D. Pye & M.J.F. Gales (1996) Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression. *Proc. ICSLP'96*, pp. 1133-1136, Philadelphia.

[43] P.C. Woodland & D. Povey (2000). Large Scale Discriminative Training for Speech Recognition. *Proc. ISCA ITRW ASR2000*, pp. 7-16, Paris.

[44] T. Zeppenfeld, M. Finke, K. Ries, M. Westphal & A. Waibel (1997). Recognition of Conversational Telephone Speech using The JANUS Speech Engine. *Proc. ICASSP'97*, pp. 1815-1818, Munich.