RAPID SPEECH RECOGNIZER ADAPTATION TO NEW SPEAKERS

Vassilis Digalakis (1) Sid Berkowitz (2) Enrico Bocchieri (3) Costas Boulis (1) William Byrne (4)
Heather Collier (5) Adrian Corduneanu (6) Ashvin Kannan (7) Sanjeev Khudanpur (4) Ananth Sankar (8)

(1) Technical U. of Crete (2) Department of Defense (3) AT&T (4) CLSP, Johns Hopkins U.
(5) West Virginia U. (6) U. of Toronto (7) Nuance Communications (8) STAR Lab, SRI

ABSTRACT

This paper summarizes the work of the "Rapid Speech Recognizer Adaptation" team in the workshop held at Johns Hopkins University in the summer of 1998. The project addressed the modeling of dependencies between units of speech with the goal of making more effective use of small amounts of data for speaker adaptation. A variety of methods were investigated and their effectiveness in a rapid adaptation task defined on the SWITCHBOARD conversational speech corpus is reported.

1. INTRODUCTION

Humans have little difficulty recognizing speech in noisy environments, speech distorted by having passed through an unknown channel or speech from nonnative speakers. We adapt to the characteristics of the new speech, often after hearing only a few seconds of it and a subset of the sounds, by exploiting relationships between various sounds. In this project, the participants alleviated the commonly used remedy of tying, or forcing to be identical, the transformations of the models of related speech units. The team followed an alternative approach, modeling the dependencies between the speech units, so that the model transformation for one unit influences but is not necessarily identical to the transformation for another unit. We used this knowledge to transform each model individually without requiring a large sample of each speech segment for adaptation. To estimate the dependencies between different speech units, we used a large corpus of training speakers and a variety of correlation modeling techniques that included Markov Random Fields, explicit correlation models, and tree-structured models.

Constrained estimation of hidden Markov models (HMMs) is currently the method of choice for adaptation by most researchers and is briefly reviewed in Section 2. In Section 3 we describe the techniques we explored in the project for modeling the dependencies between parameters of the transformation biases, and in Section 4 we present two approaches for rapid adaptation using more general transformations. Experimental results on the Switchboard corpus are presented in Section 5, and we conclude in Section 6.

2. CONSTRAINED-ESTIMATION ADAPTATION

A family of adaptation algorithms [3, 6, 8] for continuous mixture density HMMs is based on constrained reestimation of the mixture Gaussians. Maximum likelihood (ML) reestimation of the Gaussians in all these adaptation schemes is performed using the expectation-maximization (EM) algorithm.

The observation densities of the speaker-independent (SI) HMMs have the form $P_{SI}(x_t|s_t) = \sum_{i=1}^{N_\omega} p(\omega_i|s_t) N(x_t; \mu_{s_t i}, \Sigma_{s_t i})$, where x_t is the observed feature vector at time t, s_t is the HMM state at time t, ω_i denotes the event that the i-th Gaussian mixture of state s_t was used at time t, and N_ω is the number of component Gaussians in the mixture density. $N(x_t; \mu_{s_t i}, \Sigma_{s_t i})$ is the multivariate normal density with mean vector $\mu_{s_t i}$ and covariance matrix $\Sigma_{s_t i}$.

Examples of estimation constraints are that the speaker-adapted (SA) means and covariances [3], or simply the means in the *MLLR* method [6] are obtained from the SI ones through an affine transformation. The transformations are shared among states that are clustered together based on their similarity, as specified by the index $g = \gamma(s_t)$. In our work we used several special cases of the following general model for constrained-estimation adaptation

$$P_{SA}(x_t|s_t) = \sum_{i=1}^{N_\omega} \mathrm{p}(\omega_i|s_t) N(x_t; A_g \mu_{s_t i} + b_{g'}, \Sigma_{s_t i}). \quad (1)$$

By forcing $g = g'$ we obtain the usual MLLR adaptation. The number of parameters that must be estimated can be reduced by using a *structured* transform, in which the parts of the mean vectors corresponding to the cepstrum and its derivatives are transformed independently using a block-diagonal matrix. A simpler constraint is the simple *bias* transform, obtained by forcing $A_g = I$, the identity matrix. This type of transform is not as powerful as MLLR, but it is easier to model dependencies of simple biases, as we shall see in Section 3. To retain the modeling capability of the affine transform, and at the same time model dependencies between the transforms of different units, we developed in the workshop a *cascade* transformation, where the transformation components with more parameters are tied more aggressively and the clusters g' are a further refinement of the clusters g.

3. DEPENDENCY MODELING OF BIAS TRANSFORMS

In rapid adaptation, where a very small amount of speech is available for adaptation, many of the classes that share the same transformation will be either unseen or have insufficient data to robustly estimate their transformation. To predict the transformations of missing classes, as well as smooth the transformations of classes with insufficient data, we use dependency modeling. In this approach, the dependencies between the parameters of different classes are first estimated from a large number of training speakers. During adaptation, the transformation parameters of the classes that are seen in the adaptation data are estimated using standard ML techniques. These parameters are then smoothed, and the missing-class parameters are predicted from the seen ones using the dependency model that was estimated from the training speakers. Finally, the interpolated and smoothed parameters are used to transform the HMMs.

In the workshop we experimented with a number of different dependency models, including Tree-structured models, explicit correlation schemes and Markov Random fields that are explained in the remainder of this section.

3.1. Tree-structured Correlation Methods

Multiscale Tree Models: Multiscale stochastic processes based on scale-recursive dynamics on trees [1, 7, 5] are a generalization of linear dynamical systems that evolve on a tree, rather than in time. These models allow efficient algorithms for both estimation and likelihood calculation resulting in a variety of applications. A tree is defined with the biases at the leaves and a multiscale process on the tree specifies the joint distribution of the leaves.

Given a tree topology, the process parameters, and the set of all available observations (ML estimates and errors for the observed biases computed at each leaf independently), MMSE smoothed estimates of the bias and associated error covariance can be computed in a recursive but non-iterative manner [1]. The smoothed bias estimates at the leaves are then used to adapt the models within that class. Given a tree topology and observations from training data, the parameters of the tree can be estimated using an EM algorithm [5].

Tree Structured MAP Adaptation: Another tree-based Bayesian approach to adaptation, called structural MAP (SMAP) was presented in [10] and we implemented this approach for biases as a comparative exercise due to its similarity with the multiscale model. In the upward sweep of the SMAP, estimates of all parents are computed by aggregating ML estimates of the biases from the leaves. Once the aggregate bias at a parent is estimated, it serves as the prior for its immediate children. Details of the formalism are available in [10].

Qualitatively, the SMAP methodology *imposes* (or permits the designer to hand-craft) the dependence behavior between the biases through the choice of some hyperparameters, while the multiscale approach estimates the dependence structure from training data. On a quantitative note, the upward sweep of the SMAP is the same as the upward sweep of the multiscale model for a noiseless evolution with $A = I$. The downward pass of SMAP combines the parent's estimate with the child's ML estimate using an ad-hoc procedure while the downward pass of the multiscale model does an optimal fusion of information (MMSE for Gaussians) [10, 1].

3.2. Explicit Correlation Methods

The novelty here is to use a prediction model of the biases of the cascade system (1), thus combining $MLLR$ adaptation and correlation modeling, on the hypothesis that the residual error of the transformed Gaussian means are correlated. The statistics of corresponding component pairs in different bias vectors are assumed bivariate Gaussians, and estimated from the training data. During adaptation, we use linear regression to predict the bias components b_k from observed component b_j: $\hat{b}_{k/j} \;=\; \alpha_{k,j}\; b_j \;+\; \beta_{k,j}$. The predictors $\hat{b}_{k/j}$, such that the correlation $\rho_{k,j}$ exceeds a given threshold, are then interpolated according to ML estimation:

$$\hat{\hat{b}}_k \quad = \quad \sum_j w_j \hat{b}_{k/j} \;\;,\quad w_j \propto \frac{1.0}{\sigma^2_{\hat{b}_{k}/b_j}} \tag{2}$$

$$\hat{\hat{\hat{b}}}_k \quad = \quad S\, b_k \;+\; (1-S)\; \hat{\hat{b}}_k . \tag{3}$$

where $\sigma^2_{\hat{b}_k/b_j}$ is the variance of estimator $\hat{b}_{k/j}$. We use (2) both to estimate the biases unseen in the adaptation data, and to smooth the estimated biases b_k (seen data), as in (3).

3.3. Markov Random Fields

Markov Random Fields (MRFs) were first used in modeling dependencies of the Gaussian means in [9]. Here, we used MRFs to model dependencies between the biases of the cascade and simple-bias transformations. Despite the elegant theory behind MRFs, their application to correlation modeling for adaptation boils down to an implementation that is very similar to the explicit correlation techniques. The main difference is that smoothing of seen biases and prediction of unseen biases is done jointly in an iterative fashion, where the current estimates of the biases are used to obtain new estimates for both the seen and unseen classes. MRFs, SMAP and multiscale trees are all Bayesian schemes, in the broad sense of the term, and the main difference is in the form of the prior distributions. Moreover, it has been shown that MRFs and multiscale trees are equivalent [7].

4. ROBUST ESTIMATION OF TRANSFORM MATRICES

4.1. Adaptation Using Basis-Transforms

The number of transform parameters to be estimated in the MLLR adaptation method is $9D^2 + 3D$, where D is the dimensionality of the cepstrum feature vector. To address the difficulty of estimating so many parameters, we are studying a method based on "basis transforms". Here we assume a set of N basis transforms that have been estimated from a large amount of training data. Each speaker S is represented by a set of N weights $\lambda_i^S, i = 1, \ldots, N$, that are used to appropriately combine these basis transforms to create a speaker-specific model. Thus we only need to estimate on the order of 10 to 50 weights for each test speaker.

The basis transforms can be combined in two ways. *Transform combination* has been described in [4]. In another approach, which we call *density combination*, the state conditional densities induced by each basis transform are linearly combined to produce the densities for the test speaker. This is similar in spirit to ML Stochastic Transforms [2]. Evaluation of the density-combination approach on the workshop task is in progress.

4.2. Discounted-Likelihood Estimation of Transforms

The techniques discussed thus far employ constrained transformations because they require relatively little adaptation data. If more powerful transforms could be found from the same amount of data, they might perform better than these simple transformations. A method is proposed to perform iterative MLLR robustly. Iterative MLLR yields multiple transforms by iteratively estimating transforms and refining the regression classes; this is more effective than training a large numbers of transforms from scratch. The intent is to modify this procedure so that the regression class statistics can be found robustly. For this to be effective, the technique must find statistics so that they remain consistent with the transforms found at previous MLLR iterates; otherwise the iterative training procedure will be disrupted.

A variant of the EM algorithm that maximizes a *discounted likelihood criterion* [11] is applied to this problem. The modification derives from a *confidence* scale factor $c \in [0, 1]$ that is incorporated into the likelihood criterion. The M-Step is unchanged, but the statistics found by the E-Step are interpolated with those used by the M-Step at the previous iteration. The effect is to slow the convergence of the training procedure if there is insufficient data; convergence is discussed in [12].

For robust iterative MLLR, a refinement of the regression classes can be performed at each iteration so statistics found for each class are interpolated with those of its previously found parent class [1]; if the weight $c \ll 1$, the children will have transforms identical to those of their parents. There are many ways in which hierarchical MLLR transforms could be found under this algorithm. Two examples are given which do not employ reestimation of the statistics; exact implementations are being studied.

Suppose a global MLLR transform $W^{(1)}$ (incorporating rotation and shift) has been found. Using this global transform, adapted means μ and new statistics γ can be found. Suppose a transform $W^{(2)}$ is to be found for a class of states C. One possibility is to modify the MLLR reestimation criterion [6]:

$$\begin{gathered} \sum_s K_1(s)\Sigma_s^{-1}\,\gamma_s\tilde{\mu}_s\hat{\mu}_s' = \sum_s K_1(s)\;\gamma_s\;\Sigma_s^{-1}W^{(2)}\hat{\mu}_s\hat{\mu}_s' \\ K_1(s) = c_1\;1_C(s) + 1 - c_1 \end{gathered} \tag{4}$$

where $\tilde{\mu}$ is the usual EM mean reestimate and $\hat{\mu} = [1\;\;\mu]'$. For $c_1 = 0$, all states contribute to the update so that $W^{(2)}$ is a two-iteration estimate of the global transform. If $c_1 = 1$, $W^{(2)}$ is a class-specific transform initialized by the global transform. For other values of c_1, $W^{(2)}$ is something between the two. In effect, c_1 specifies the confidence with which a unique transform should be estimated for class C.

Consider next a set of states $C' \subset C$ for which a transform $W^{(3)}$ is to be estimated in a 'smooth manner' from $W^{(2)}$. Let μ be the means transformed by $W^{(2)}$. $W^{(3)}$ may be found to satisfy (4), where $K_1(s)$ is replaced by $K_2(s) = c_2\;\;1_{C'}(s) + (1-c_2)K_1(s)$. For $c_2 = 0$, this criterion becomes (4) so that $W^{(3)}$ is (nearly) an identity transform. The reestimate of $W^{(3)}$ can therefore back-off gradually to the robustly estimated parent transform $W^{(2)}$.

5. EVALUATION ON THE SWITCHBOARD CORPUS

5.1. Task Definition and Baseline Results

We used the Switchboard database to evaluate the various adaptation methods. *Transcription-mode* adaptation has been previously applied to this task, by adapting the recognizer to the same data that it is being tested. This, of course, is done in unsupervised mode (i.e. without using knowledge about what has been said). In our work, however, we wanted to be able to evaluate in both unsupervised and supervised modes, since the error rates in Switchboard high. When these two benchmarks are different, supervised mode allows us to better evaluate the adaptation-rate characteristics of an algorithm. Supervised performance can be evaluated in *batch-mode* adaptation, where we adapt using speech different from the test data.

We evaluated rapid adaptation using two batch-mode benchmarks, with 30 and 60 seconds of speech, respectively. The two benchmarks were defined on the 1997 summer-workshop development set by equally splitting the speech of each conversation side into two parts, adapting on the first 30 or 60 seconds of each part, and testing on the other half. The complete definition of the task can be found on the 1998 Workshop web site [13]. The SI system [2] was a speaker- and gender-independent context-dependent HMM system. The speaker-independent performance of this system on the development set was 45.3%. After optimizing the number of transformations, the cascade-transformation outperformed slightly the standard block-diagonal MLLR with word-error rates of 42.6% (30" unsupervised), 42.0% (30" supervised), 41.4% (60" unsupervised) and 40.2% (60" supervised).

In addition, we evaluated the adaptation performance of three cases: two simple-bias transforms with 150 and 250 classes, respectively, and one for a cascaded transform with one global MLLR transformation 150 classes for the biases. These configurations were used in the dependency-modeling experiments described below. The corresponding recognition results using the ML estimates of the biases for the four different benchmarks are summarized in Table 1.

5.2. Tree-structured Correlation Methods

Multiscale Tree Models: Multiscale tree models of dependence are presented for three cases: two to estimate biases and one to estimate cascaded biases in conjunction with a global MLLR transform. The Gaussian densities in the system were divided into either 150 (or 250) classes and ML estimates of the class-biases or cascaded class-biases were obtained for the nearly 3000 conversation sides in the acoustic training corpus. These were used to train dependence models for the different systems For the test speakers, ML estimates of the corresponding biases were obtained for those classes which had sufficient data, and the multiscale tree models were used to obtain smoothed estimates of all the 150 (250) class biases. Table 1 shows the results of smoothing the biases or cascaded-bias transforms using multiscale tree models. The performance of the ML estimates of the corresponding biases (before smoothing) are also shown for comparison. The WER of the unadapted system is 45.2%. It is interesting to note that even with less data (30 sec) smoothing of the 250 biases is as good as that of 150 - in accordance with the multiscale theory.

Mode and Dur (sec.)	150-bias		250-bias		150-casc.	
	ML	MS	ML	MS	ML	MS
Sup (30)	44.3	43.2	44.9	43.0	41.7	41.2
Unsup (30)	45.0	43.8	45.5	44.1	43.2	42.3
Sup (60)	42.9	42.1	43.0	42.1	40.2	40.0
Unsup (60)	44.6	43.6	44.5	43.2	42.2	41.8

Table 1: Recognition WER (%) for smoothing ML estimates of class-biases using multiscale tree models (MS)

Tree Structured MAP Adaptation: Recognition results are presented for the SMAP scheme with three different settings of the hyperparameters (τ). In the first two cases, we used a constant value of τ (1 and 10) for each level in the tree, and in the third case τ increased from 0.3 at the root to 10 at the leaves, indicating high faith in the ML estimate of the aggregate (class) bias at the root and relatively less confidence in the detailed ML estimate of the bias of the individual Gaussian components. Table 2 shows the performance of the SMAP technique. The variable-τ case is indicated by $\tau = 0.3$. The performance of ML bias adaptation

[1] M. Gales reports that similar Bayesian formulations exist.

[2] The system was trained with 60 hours of per-utterance cepstral-mean normalized speech, and the adaptation experiments were done by rescoring lattices created with a 22,000 bigram language model.

and that of the multiscale tree model of biases with 250 classes is shown alongside. [3]

Mode and Dur (sec.)	250 ML	SMAP with $\tau =$ 1.0	10.0	0.3	250 MS
Sup (30)	44.9	44.0	43.8	43.2	43.0
Unsup (30)	45.5	45.4	44.8	44.5	44.1
Sup (60)	43.0	43.3	43.2	42.2	42.1
Unsup (60)	44.5	45.5	44.7	44.5	43.2

Table 2: Recognition WER (%) for 250-class ML biases, SMAP adaptation and 250-class multiscale tree models (MS)

5.3. Explicit Correlation Methods

We have applied the correlation model (2 to the cascade system (1) with one transform and 150 biases.The results in Table 3 show improvements ranging from 0.3% to 0.9%.

In addition, we evaluated the MRF dependency model for the bias-transformation systems. The model improved the unsmoothed biases in all cases, and the results were similar to the multiscale-tree smoothing results shown in Table 1. Specifically, the 30" and 60" unsupervised benchmarks for the 150-bias configuration gave a WER of 43.6% and 43.7%, respectively, whereas the corresponding results for the 250-bias system were 43.7% and 43.3%.

Adapt. Mode	Adapt. Data	*Weight S in (2)* 0.8	0.7	0.6	0.5	0.4
Unsup.	30"	42.5	42.3	42.3	42.4	42.5
	60"	41.8	41.8	41.8	41.9	42.1
Sup.	30"	41.5	41.4	41.4	41.3	41.3
	60"	40.0	39.9	40.0	40.1	40.4

Table 3: Correlation modeling of cascade biases recognition WER.

5.4. Discounted Likelihood Methods

The techniques for the robust estimation of multiple regression class MLLR transforms proposed in Section 4.2 have been evaluted on the unsupervised 60" adaptation task. The baseline 4 MLLR transforms yield a WER of 42.1%. Using the first technique of Section 4.2, 4 transforms can be found from the global transform that yield an improved WER of 41.6%. Further gains are possible by estimating 11 and 20 transforms from these 4 transforms, yielding 1.0% over the baseline 4 transform performance.

Num. Transforms	1	4	11	20
Baseline	42.7	42.1	42.2	
$c_1 = 0.5$		42.3		
$c_1 = c_2 = 0.9$		41.6	41.1	41.3

Table 4: WER of Discounted Likelihood Full Transforms.

For $c_1 = 0.5$, the 4 transform system is relaxing towards the performance of the single transform system; at $c_1 = 0.0$ their performance would be identical. Note that 11 transforms cannot be well-estimated in the usual manner on this 60" task (the quoted result given is from a block diagnal system), whereas even 20 transforms can be fairly robustly estimated using the newly proposed method.

[3] Due to limitations in available software our comparison of the two schemes is not quite exact. The SMAP scheme was implemented componentwise in the bias vectors, while the multiscale model has been implemented with full error covariances estimated across training speakers instead of using true error covariances of the biases.

6. CONCLUSIONS

Detailed studies of the effectiveness of dependency models for rapid adaptation are reported. Effective techniques are available and their effectiveness depends upon the power of the transform they model. A detailed analysis [13] of ASR performance of these systems indicates that (i) systems improve almost equally on seen and unseen words on supervised adaptation, as expected; (ii) in unsupervised adaption, unseen words also improve, but among seen words, only correctly adapted words improve.

ACKNOWLEDGMENTS This material is based upon work supported by the National Science Foundation under Grant No. #IIS-9732388, and carried out at the 1998 Workshop on Language Engineering, Center for Language and Speech Processing, Johns Hopkins University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or The Johns Hopkins University.

7. REFERENCES

[1] K. Chou, A. Willsky, and A. Benveniste, "Multiscale recursive estimation, data fusion, and regularization," *IEEE Trans. Automatic Control*, Mar. 1994.

[2] V. Diakoloukas and V. Digalakis, "Adaptation of Hidden Markov Models Using Multiple Stochastic Transformations," *Proc. European Conf. on Speech Comm. and Tech.,* Sept. 1997.

[3] V. Digalakis, D. Rtischev and L. Neumeyer, "Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures," *IEEE Trans. Speech and Audio Processing,* Sept. 1995.

[4] M. Gales, "Cluster Adaptive Training For Speech Recognition," *Proc. International Conf. on Spoken Language Processing,* 1998, to appear.

[5] A. Kannan and M. Ostendorf, "Modeling dependence in adaptation of acoustic models using multiscale tree processes," in *Proc. European Conf. on Speech Comm. and Tech.*, 1997.

[6] C. Leggetter and P. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language,* 1995.

[7] M. Luettgen, W. Karl, A. Willsky, and R. Tenney, "Multiscale representations of Markov random fields," *IEEE Trans. on Signal Proc.*, v. 41, Dec. 1993.

[8] A. Sankar and C.-H. Lee, "A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Trans. Speech and Audio Processing,* May 1996.

[9] B. Shahshahani, "A Markov Random Field Approach to Bayesian Speaker Adaptation," *IEEE Trans. Speech and Audio Processing,*

[10] K. Shinoda and C-H. Lee, "Structural MAP speaker adaptation using hierarchical priors," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

[11] W. Byrne, "Generalization and Maximum Likelihood from Small Data Sets," *Proc. IEEE-SP Workshop on Neural Networks for Signal Processing*, 1993.

[12] W. Byrne, A. Gunawardana, and S. Khudanpur, "Information Geometry and EM Variants," submitted to *IEEE Trans. on Signal Processing*.

[13] *http://www.clsp.jhu.edu/ws98/projects/adapt*.