CORRELATION MODELING OF MLLR TRANSFORM BIASES FOR RAPID HMM ADAPTATION TO NEW SPEAKERS

Enrico Bocchieri (1)   Vassilis Digalakis (2)   Adrian Corduneanu (3)   Costas Boulis (2)

(1) AT&T Research   (2) Technical U. of Crete   (3) U. of Toronto

ABSTRACT

This paper concerns *rapid* adaptation of hidden Markov model (*HMM*) based speech recognizers to a new speaker, when only few speech samples (one minute or less) are available from the new speaker. A widely used family of adaptation algorithms defines adaptation as a linearly constrained reestimation of the *HMM* Gaussians. With few speech data, tight constraints must be introduced, by reducing the number of linear transforms and by specifying certain transform structures (e.g. block diagonal). We hypothesize that under these adaptation conditions, the residual errors of the adapted Gaussian parameters can be represented and corrected by dependency models, as estimated from a training corpus. Thus, after introducing a particular class of linear transforms, we develop correlation models of the transform parameters. In *rapid* adaptation experiments on the *SWITCHBOARD* corpus, the proposed algorithm performs better than the transform-constrained adaptation and the adaptation by correlation modeling of the *HMM* parameters, respectively.

1. INTRODUCTION

Adaptation techniques have been developed for automatic speech recognizers to compensate for differences between the speech on which the system was trained, and the speech which it has to recognize. However, before obtaining significant improvement in recognition performance, several minutes of speech from the new speaker or environment must be provided. *Rapid* adaptation, where the recognizer has to adapt on one minute or less of speech data, is not as effective. On the contrary, humans can quickly adapt to the characteristics of speech distorted by an unknown channel, or pronounced by a nonnative speaker. Humans seem to exploit relationships between various speech sounds, so that, having heard only few speech samples in a new distorted environment, they adjust to all the speech pronounced in this environment. Therefore, in the speaker adaptation project of the 1998 workshop organized by the CLSP center at Johns Hopkins University, many participants decided to improve rapid speech recognizer adaptation by modeling the dependencies between speech sounds (see [1] for project summary). In particular, this paper provides specific details about one of the adaptation approaches based on dependency modeling, developed at the workshop.

Dependency models for the speaker adaptation problem have been long introduced [2, 3] and further studied by several other authors [4, 5, 6, 7, 8, 9]. However, the most widely used hidden Markov model (*HMM*) adaptation algorithms, reviewed in Section 2, take a different approach. Adaptation is implemented as reestimation of the *HMM* under certain constraints, defined by linear transforms of the *HMM* parameters. With few adaptation data, tight constraints must be introduced by *tying* linear transforms in a small number of classes and by specifying certain transform structures (e.g. block diagonal). Our hypothesis is that, under these adaptation conditions, the residual errors of the adapted *HMM* parameters can be represented and corrected by correlation models, as estimated from a large training corpus. In this study, to take advantage of the correlation between the transform parameters, we introduce a particular class (*cascade*) of linear transforms (Section 3). We use correlation models to predict the values of transform parameters that cannot be estimated from the adaptation data, and to improve the parameter estimates by smoothing (Section 4). Finally, speaker adaptation experiments on the *SWITCHBOARD* corpus are detailed in Section 5, and conclusions drawn in Section 6.

2. MLLR ADAPTATION

A family of adaptation algorithms [10, 11, 12] for continuous density *HMM* is based on constrained reestimation of the mixture Gaussians. Maximum likelihood (*ML*) reestimation of the Gaussians is based on the expectation-maximization (*EM*) algorithm [13]. Let the observation densities of the speaker independent (SI) *HMM* be Gaussian mixtures:

$$P_{SI}(\mathbf{x_t}|s_t) = \sum_{i=1}^{N_\omega} p(\omega_i|s_t) N(\mathbf{x_t}; \boldsymbol{\mu}_{s_t,i}, \Sigma_{s_t i}), \qquad (1)$$

where $\mathbf{x_t}$ is the observed feature vector at time t, s_t is the *HMM* state, 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. $N(\mathbf{x_t}; \boldsymbol{\mu}_{s_t,i}, \Sigma_{s_t i})$ is the multivariate normal density with mean $\boldsymbol{\mu}_{s_t,i}$ and covariance $\Sigma_{s_t i}$.

In the *MLLR* algorithm [11], the linear constraint is applied to the means of the adapted observation densities. The adapted state Gaussian mixture becomes:

$$P_{SA}(\mathbf{x_t}|s_t) = \sum_{i=1}^{N_\omega} \mathrm{p}(\omega_i|s_t) N(\mathbf{x_t}; A_g \boldsymbol{\mu}_{s_t,i} + \mathbf{b_g}, \Sigma_{s_t i}). \qquad (2)$$

The transformations are shared among states according to similarity of states, as specified by the index $g = \gamma(s_t)$. To reduce the number of parameters that must be estimated, a *structured* transform is often used [14], by transforming independently the cepstrum and the cepstrum derivative components with a block-diagonal matrix. A simpler constraint is the *bias* transform, that implements adaptation as an additive Gaussian mean bias:

$$P_{SA}(\mathbf{x_t}|s_t) = \sum_{i=1}^{N_\omega} \mathrm{p}(\omega_i|s_t) N(\mathbf{x_t}; \boldsymbol{\mu}_{s_t,i} + \mathbf{b_g}, \Sigma_{s_t i}). \qquad (3)$$

(3) is not as powerful as the affine transformation in (2). However, it is easier to model dependencies of simple biases.

3. CASCADE MLLR TRANSFORMS

To retain the modeling capability of the affine transform, and to model dependencies between transforms of different speech units, we have defined *cascade* transforms, by tying less aggressively the transform components with fewer parameters:

$$P_{SA}(\mathbf{x_t}|s_t) = \sum_{i=1}^{N_\omega} \mathrm{p}(\omega_i|s_t) N(\mathbf{x_t}; A_g \boldsymbol{\mu}_{s_t,i} + \mathbf{b_{g'}}, \Sigma_{s_t i}). \quad (4)$$

The clusters g', used in (4) for tying biases $\mathbf{b_{g'}}$, are a refinement of g. The term *cascade* refers to the adopted estimation algorithm of $\mathbf{b_{g'}}$. First, the *MLLR* transforms (2) are estimated. Then, an additional *EM* step on cluster g' refines the bias estimate, yielding $\mathbf{b_{g'}}$.

When $\mathbf{b_{g'}}$ cannot be estimated (unseen adaptation data), the less refined $\mathbf{b_g}$ provides a *backoff* estimate.

4. CORRELATION METHODS FOR THE CASCADE TRANSFORM

With few adaptation data available, many of the biases $\mathbf{b_{g'}}$ (4) cannot be estimated (unseen data), or the estimates (seen data) may be unreliable. Rather than backing-off, we use correlation models to predict the unseen biases and to smooth the seen biases.

4.1. Explicit Correlation Model

We assume that the statistics of corresponding component pairs in different *cascade* bias vectors (4) are bivariate Gaussians. These are estimated prior to adaptation, from a large corpus of training data, collected by many speakers (different from the adaptation subjects). In adaptation, we predict the components of $\mathbf{b_{k'}}$ from the components of seen bias $\mathbf{b_{j'}}$ through linear regression: $\hat{b}_{k'/j'} = \alpha_{k',j'}\, b'_j + \beta_{k',j'}$. All the predictors $\hat{b}_{k'/j'}$, such that the correlation $\rho_{k',j'}$ exceeds a threshold, are interpolated according to maximum likelihood:

$$\hat{\hat{b}}_{k'} = \sum_{j'} w_{j'} \hat{b}_{k'/j'} \;, \quad w_{j'} \propto \frac{1.0}{{\sigma_{\hat{b}_{k'}/b_{j'}}}^2} \quad (5)$$

where

$${\sigma_{\hat{b}_{k'}/b_{j'}}}^2 = {\sigma_{b_{k'}}}^2 (1 - \rho^2_{b_{k'},b_{j'}}) + {\alpha_{k',j'}}^2\, {\sigma^2}_{\delta_{b'_j}} \quad (6)$$

is the variance of estimator $\hat{b}_{k'/j'}$, and ${\sigma^2}_{\delta_{b'_j}}$ is the variance of the predictor source estimate [4].

We use (5) to predict unseen bias $\mathbf{b_{k'}}$, eventually after smoothing with the backoff value $\mathbf{b_k}$ of $\mathbf{b_{k'}}$:

$$\hat{\hat{\hat{\mathbf{b}}}}_{k'} = W\, \hat{\hat{\mathbf{b}}}_{k'} + (1 - W)\, \mathbf{b_k} \quad (7)$$

and to smooth the estimates $\mathbf{b_{k'}}$ (seen data):

$$\hat{\hat{\hat{\mathbf{b}}}}_{k'} = S\, \mathbf{b_{k'}} + (1 - S)\, \hat{\hat{\mathbf{b}}}_{k'} \quad (8)$$

4.2. Markov Random Fields

Markov random fields (*MRF*s) were first used in modeling dependencies of the Gaussian means in [6]. We use *MRF*s to model dependencies between the biases of the *cascade* and simple bias transformations. Despite the elegant theory behind *MRF*s, their application to correlation modeling for adaptation gives an implementation that is very similar to the explicit correlation technique. 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.

In our implementation, the new estimate for the bias element of a class is given by

$$b_s^{new} = \frac{w_{s,s} b_s^{old} + \sum_{r \in N_s} w_{s,r} (\alpha_{s,r} b_r^{old} + \beta_{s,r})}{w_{s,s} + \sum_{r \in N_s} w_{s,r}} \quad (9)$$

where the prediction neighbors (the set N_s) for a particular point s are these elements r of class biases that are mostly correlated to it. The coefficients $\alpha_{s,r}, \beta_{s,r}$ are obtained through linear regression, whereas the combination weights are based on the variance of the MAP estimates,

$$w_{s,r} = \begin{cases} \dfrac{N_s + \tau}{\eta}, & s = r \\ \dfrac{1}{1 - \rho^2_{b_s,b_r} (1 - \frac{\eta}{N_r + \tau})}, & s \neq r \end{cases} \quad (10)$$

where N_s is the number of observations that were used to estimate the bias s and τ, η are constants determined empirically.

5. RESULTS ON THE SWITCHBOARD CORPUS

5.1. Task Definition

We used the *SWITCHBOARD* speech corpus to test the adaptation methods. Unsupervised *transcription-mode* adaptation has been previously applied to this task, by adapting the recognizer to the same data that is being recognized (details in [1, 15]). We evaluated *rapid* adaptation on two batch-mode benchmarks, with 30 and 60 seconds of speech, respectively. The 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 [15]. The speaker and gender independent *HMM* was trained with 60 hours of speech data, which were also used for dependency model estimation. We used per-utterance cepstral-mean normalization, and we tested adaptation performance by rescoring lattices with a 22,000 bigram language model.

5.2. Baseline And Cascade MLLR

The speaker-independent word error rate on the development set was 45.3%. The baseline *MLLR* adaptation (2) was tested with various numbers of transforms and transform structures. The transform classes were chosen according to acoustic phonetic knowledge. The system with four block diagonal transforms gave the best results (Table 1), both with 30 and 60 seconds of adaptation data.

We optimized the cascade adaptation (4) by experiments with different number of matrices A_g and biases $\mathbf{b_{g'}}$. The best results were obtained with one matrix and either 11 or 21 biases, depending on the amount of adaptation data. Table 2 also shows the results with 150 *cascade* biases (one for every monophone state),

Adapt Mode	Transform. Type	Adapt. Data	Number of Transforms	Word Error
SI	-	-	-	45.3%
Unsup.	Full	30"	1	43.1%
	Full		4	44.5%
	Block		1	42.9%
	Block		4	42.6%
	Block		11	42.8%
	Full	60"	1	42.7%
	Full		4	42.1%
	Block		1	42.8%
	Block		4	42.1%
	Block		11	42.2%

Table 1: Speaker-independent and *MLLR* adaptation results.

Adapt. Mode	Transform Type	Adapt. Data	Number of Transf./Biases	Word Error
Unsup.	Block	30"	4/4	42.6%
	Cascade		1/11	42.6%
	Cascade		1/150	43.2%
	Block	60"	4/4	42.1%
	Cascade		1/21	42.0%
	Cascade		1/150	42.2%
Sup.	Block	30"	4/4	41.6%
	Cascade		1/11	41.4%
	Cascade		1/150	41.7%
	Block		4/4	40.8%
	Cascade	60"	1/21	40.8%
	Cascade	60"	1/150	40.2%

Table 2: Adaptation results for block-diagonal and cascade transformations.

and with the best baseline block MLLR system. The cascade-transformation outperformed slightly the standard block-diagonal MLLR.

5.3. Correlation Models For Cascade Biases

As shown in Table 2, the *cascade* systems with 11 and 21 biases are the most accurate. However, they typically give small correlation estimates between components of different bias vectors, because the biases are much smoothed by tying into relatively small number of classes. The correlations are much stronger for the *cascade* system with 150 biases. Therefore we have applied the correlation modeling of Section 4.1 to the 150 bias system, even if this is not the best of Table 2. In the *cascade* adaptation, the refined biases $\mathbf{b}_{\mathbf{g}'}$ (4) are estimated if the number of adaptation samples is greater than a given threshold (20 gave the best result). On average, only 49 and 78 of the 150 biases are estimated from 30 and 60 seconds of adaptation data, respectively.

We predict the unseen biases as per equation (7), with the maximum likelihood weights $w_{j'}$ (5). We have experimented also with predictors based only on the most correlated source and with equal weight predictors. The maximum likelihood weights give the best results (see Table 3), especially if the predictor is not smoothed with the backoff bias ($W = 1.0$).

Adapt. Mode	Adapt. Data	Predictor weights	W of (7) 0.5	1.0
Unsup.	30"	a: most correlated only	42.9%	43.1%
		b: equal weights	42.7%	42.8%
		c: max. likelihood	42.7%	42.7%

Table 3: Word error rates (%) for prediction of unseen *cascade* biases by explicit correlation model.

Adapt. Mode	Adapt. Data	*Weight S in (8)* 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 4: Word error rates (%) for prediction and smoothing of *cascade* biases by explicit correlation model.

As shown in Table 4, smoothing the estimated (seen) biases with the prediction model (8) gives further improvements. We asses the overall improvements of the correlation model of the *cascade MLLR* biases, by comparing Table 4 with the baseline system (4 block diagonal transforms) in Table 2. On the different adaptation tasks, word error rate improvements range from 0.3% to 0.9%.

5.4. Correlation Models For Gaussian Mean Biases

We now verify that correlation modeling of the *cascade* biases (4), evaluated in Section 5.3, is more effective than correlation modeling of the Gaussian mean biases (3). At the workshop, we have applied the correlation models of Section 4 to the Gaussian mean biases (3). *MRF*s with correlation smoothing gave the best results, shown in Table 5 (details in [15]).

The comparison of Tables 5 and 4, demonstrates that correlation modeling of the *cascade* biases is more effective than correlation modeling of the Gaussian mean biases.

Adapt. Mode	Adapt. Data	Number of Biases (3)	Word Error
Unsupervised	30"	150	43.6%
		250	43.7%
	60"	150	43.7%
		250	43.3%

Table 5: Word error rates (%) for *MRF* model of Gaussian mean biases (3).

6. CONCLUSIONS

In essence, the proposed algorithm for *rapid* recognizer adaptation to new speakers is effective because the *cascade* transform (4)

a) retains the modeling power of the matrix A_g, as in linearly constrained reestimation, and,

b) by relaxing the tying of the biases $\mathrm{b}_{\mathrm{g}'}$, it retains sufficient detail in the adaptation of the different speech sounds to allow for effective dependency modeling. See comments in Section 5.3 for details.

The proposed approach combines the *MLLR* and correlation based adaptation. In *rapid* adaptation experiments on the *SWITCHBOARD* corpus, the novel approach is more effective than either *MLLR* or correlation modeling used alone.

In addition, we feel that our current implementation of the explicit correlation model can be improved further. The prediction (5) is limited to corresponding components across different biases. For example the *first* cepstrum coefficient of a bias is predicted only from the *first* cepstra of other biases. This limitation should be removed, in particular because it hinders the prediction of the 1^{st} and 2^{nd} differential cepstra, that are often strongly correlated to the cepstra. We should evaluate also the *MRF* approach to correlation modeling of *cascade* biases (4), besides the explicit correlation model reported in this paper.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. (#IIS-9732388), and was 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 author(s) and do not necessarily reflect the views of the National Science Foundation or The Johns Hopkins University.

7. REFERENCES

[1] V. Digalakis et al.,"Rapid Speech Recognizer Adaptation to New Speakers," submitted to Proc. ICASSP-99, Phoenix, Arizona, 1999.

[2] Stern,R.M., and Lasry,M.J.,"Dynamic Speaker Adaptation for Feature-Based Isolated Word recognition," Trans. IEEE ASSP-35(6), pp 751-762, 1987.

[3] Cox,S.J.,"Speaker Adaptation Using A Predictive Model," Proc. of EuroSpeech, pp 2283-2286, Berlin, 1993.

[4] Ahadi,S.M., Woodland,P.C., "Combined Bayesian And Predictive Techniques For Rapid Speaker Adaptation ...," Computer Speech and Language 11, pp 187-206, 1997.

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

[6] B. Shahshahani, "A Markov Random Field Approach to Bayesian Speaker Adaptation," *IEEE Trans. Speech and Audio Processing,* pp. 183–191, March 1997.

[7] 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.

[8] Chen,S.S., DeSouza,P, "Speaker Adaptation By correlation (ABC)," Proc. Eurospeech97, Rhodes, Greece.

[9] Afify,M., Gong,Y., Haton,J. "Correlation Based Predictive Adaptation Of Hidden Markov Models," Proc. Eurospeech97, pp 2059-2062, Rhodes, Greece.

[10] V. Digalakis, D. Rtischev and L. Neumeyer, "Speaker Adaptation Using Constrained Reestimation of Gaussian Mixtures," *IEEE Transactions Speech and Audio Processing,* pp. 357–366, September 1995.

[11] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language,* pp. 171–185, 1995.

[12] A. Sankar and C.-H. Lee, "A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Transactions Speech and Audio Processing,* pp. 190–202, May 1996.

[13] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood Estimation from Incomplete Data," *Journal of the Royal Statistical Society (B)*, Vol. 39, No. 1, pp. 1–38, 1977.

[14] L.Neumeyer, A.Sankar and V.Digalakis, "A Comparative Study of Speaker Adaptation Techniques", *Proceedings of European Conference on Speech Communication and Technology*, pp. 1127–1130, Madrid, Spain, 1995.

[15] "Rapid Recognizer Adaptation," Project Web page, *http://www.clsp.jhu.edu/ws98/projects/adapt.*