LATENT VARIABLE SPEAKER ADAPTATION OF GAUSSIAN MIXTURE WEIGHTS AND MEANS

Xueru Zhang, Kris Demuynck, Hugo Van hamme*

Katholieke Universiteit Leuven, Department of Electrical Engineering - ESAT
Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven, Belgium
{Xueru.Zhang,Kris.Demuynck,Hugo.Vanhamme}@esat.kuleuven.be

ABSTRACT

We describe a novel fast speaker adaptation algorithm for large vocabulary speech recognition systems, which adapts both the Gaussian means and the mixture weights. Gaussian means are expressed as a linear combination of eigenvoices estimated with principal component analysis. The non-negative Gaussian mixture weights are expressed as a linear combination of a set of latent vectors estimated with non-negative matrix factorization. Experiments on the Wall Street Journal database show that the combination of weight and mean adaptation consistently improves the performance compared to eigenvoice adaptation only. Improvements up to 5.8% relative word error rate reduction were observed with 40 eigenvoices and 40 latent weight vectors. Furthermore, combining weight and mean adaptation outperformed both weight and mean adaptation on itself, even if the latter uses more latent vectors.

Index Terms— non-negative matrix factorization, eigenvoice and weight adaptation, speaker adaptive training, fast speaker adaptation, latent variable method

1. INTRODUCTION

The large variability in speech signals is one aspect that makes speech recognition a difficult task. A major component in this variability is the inter-speaker variation which is caused by variations in speaker characteristics such as gender, age, dialect, vocal tract length, etc. Different techniques have been proposed to compensate for this speaker variability. These techniques can be classified into two categories: feature-based transformations and model-based adaptation. The former techniques transform the feature vectors so that the derived speech feature vectors are more robust to speaker variations. Vocal tract length normalization [1] (VTLN) for example, compensates the variability resulting from physical vocal tract length differences among speakers. Model-based adaptations adjust the speaker independent (SI) acoustic models towards the speaker dependent (SD) models. By tuning the system to be speaker specific, a considerable amount of variability does not have to be modeled, which significantly improves the performance of the recognition system. Some well known and widely used model-based adaptation techniques are maximum a posteriori (MAP) adaptation [2, 3] and unconstrained and constrained (feature-space) maximum likelihood linear regression ((c)MLLR) [4]. In this paper, we focus on coping with the acoustic mismatches resulting from speaker variability by incorporating model-based speaker adaptation techniques on top of VTLN.

*This work is funded by the Dutch-Flemish IMPact program (ICTRegie-IBBT(Interdisciplinary Institute for Broadband Technology)).

When large amounts of enrollment data are available, MAP and (c)MLLR are prevalently applied. MAP maximizes the posterior probabilities of the model parameters (means and variances) with the SI model parameters as priors. It converges to the maximum likelihood (ML) estimate with large amounts of enrollment data. A drawback of MAP is that only those model parameters are updated for which sufficient data are observed. (c)MLLR applies linear transformations to the SI model parameters to generate more speaker specific acoustic models. It has a large number of parameters to be estimated and hence requires large amounts of enrollment data.

However, in many applications only limited amounts of enrollment data are available. Under these situations, rapid speaker adaptation techniques become quite appealing. In [5], the number of coefficients are reduced by expressing the cMLLR transformation matrix as a weighted sum of basis matrices. Another approach to achieve rapid speaker adaptation on small amounts of enrollment data is to constrain the adaptable model parameters to a small subspace. In eigenvoice fast speaker adaptation [6], the adapted means are expressed as a linear combination of eigenvoices. These eigenvoices are the prominent principal components of the speaker dependent mean vectors of the training speakers estimated using principal component analysis (PCA). The above listed model-based speaker adaptation techniques all focus on adapting the means (and variances) of the acoustic models. In [7] a ML-based rapid speaker adaptation algorithm which adapts the Hidden Markov Models (HMM) Gaussian mixture weights has been investigated. Similar to eigenvoice adaptation, the Gaussian mixture weights of the evaluation speakers are expressed as a linear combination of a set of latent vectors. Considering the Gaussian mixture weights are non-negative values that sum up to one, PCA (eigenvoice adaptation) cannot be applied directly. Instead, non-negative matrix factorization (NMF) [8] which performs matrix factorization under the constraints that elements of all the matrices are non-negative, is used to estimate the latent vectors.

This paper starts from the observation that both eigenvoice and weight adaptations are ML-based adaptation methods which rely on a set of latent vectors and allow rapid speaker adaptation. In this paper, we analyze and compare these two rapid speaker adaptation techniques. We further combine NMF weight-based adaptation with eigenvoice-based Gaussian mean adaptation to improve the performance of large vocabulary speech recognition systems.

This paper is organized as follows. In section 2, we give an overview of the eigenvoice speaker adaptation technique. In section 3, the NMF weight-based adaptation technique is recapitulated and the relations with eigenvoice speaker adaptation are explained. We introduce the proposed rapid speaker adaptation algorithm in section 4. In section 5 we describe our speech recognition system and

compare the recognition results obtained with the different speaker adaptation algorithms. Conclusions are presented in section 6.

2. EIGENVOICE SPEAKER ADAPTATION

Eigenvoice speaker adaptation [6] is used to perform rapid speaker adaptation by adjusting the SI Gaussian mean vector μ_{SI}. The evaluation speaker mean vector is expressed as a linear combination of eigenvoices, which are derived from the training speakers. Firstly, the SD mean vectors $\{\mu_r\}_{r=1}^{R}$ of the training speakers are estimated, with r the training speaker index and R the total number of training speakers. Next, PCA is used to estimate the eigenvectors from the covariance or correlation matrix of the training speaker SD mean vectors. The J eigenvectors with the largest corresponding eigenvalues are kept as eigenvoices $\{\phi_j\}_{j=1}^{J}$ with j the eigenvoice index. The mean vector μ_e of the evaluation speaker e can be written as a weighted combination of eigenvoices

$$\mu_e = \phi_1\rho_1 + \sum_{j=2}^{J}\phi_j\rho_j \tag{1}$$

with weighting coefficients $\{\rho_j\}_{j=2}^{J}$. The first eigenvoice ϕ_1 is set to the SI mean vector μ_{SI}, with corresponding coefficient $\rho_1 = 1.0$.

The coefficients $\{\rho_j\}_{j=2}^{J}$ are estimated by maximizing the likelihood of the enrollment data. As is described in [6], this can be done by solving matrix equation (2).

$$\sum_{s,k,t}\gamma_{e;sk}(t)\phi_{j;sk}^{T}\Sigma_{\text{SI};sk}^{-1}(\mathbf{o}_t - \mu_{\text{SI};sk}) = \sum_{s,k,t}\gamma_{e;sk}(t)\left\{\sum_{i=2}^{J}\rho_i\phi_{i;sk}^{T}\Sigma_{\text{SI};sk}^{-1}\phi_{j;sk}\right\}, \quad \forall j = 2\cdots J \tag{2}$$

with $\mathbf{o}_t$ the observation at time t, s the state index, k the Gaussian index at state s, $\gamma_{e;sk}$ the posterior probability of Gaussian k at state s of evaluation speaker e. $\mu_{\text{SI;sk}}$ and $\Sigma_{\text{SI;sk}}$ are the speaker independent mean and covariance matrix of Gaussian k at state s respectively.

The eigenvoices encode the relations between the SD means of the training speakers. By exploiting these relations, means of Gaussians not activated during the enrollment phase (unobserved Gaussians) will still be updated, by extrapolating their behavior based on those that were observed. Furthermore, the number of degrees of freedom equals the number of eigenvoices minus one ($\phi_1 = \mu_{\text{SI}}$, $\rho_1 = 1.0$). Given that the number of eigenvoices is small, only few parameters have to be estimated and hence the eigenvoice speaker adaptation technique can be used to compensate speaker variability given small amounts of enrollment data.

3. GAUSSIAN MIXTURE WEIGHT ADAPTATION

The principle of the HMM Gaussian mixture weight adaptation [7] is to adapt the weights instead of the means (and variances) of the SI acoustic models. Similar to eigenvoice speaker adaptation, the Gaussian mixture weights are expressed as a weighted sum of a set of latent vectors. These latent vectors are estimated from the SD Gaussian mixture weights of the training speaker. Considering the Gaussian mixture weights are probabilities, the estimated latent vectors are also required to contain proper probability distributions, i.e. non-negative values that sum up to 1.0. Different from eigenvoice speaker adaptation where PCA is used to estimate the eigenvoices from the mean vectors, NMF [8] is applied to estimate the non-negative latent vectors by maximizing the likelihood of the training data.

NMF approximates a non-negative matrix $\mathbf{V}$ as a product of two non-negative matrices: a latent vector matrix $\mathbf{W}$ and the latent vector coefficient matrix $\mathbf{H}$. For the NMF speaker adaptation, the matrices $\mathbf{W}$ and $\mathbf{H}$ are chosen to maximize the training data likelihood, which is equivalent to maximize the following auxiliary function:

$$Q(\mathbf{W},\mathbf{H}) = \sum_{r,s,k,t}\gamma_{r;sk}(t)\log(\lambda_{r;sk}), \tag{3}$$

with Gaussian mixture weight for training speaker r expressed as a linear combination over the L latent vectors $w_{(s,k),l}$

$$\lambda_{r;sk} = \sum_{l=1}^{L} w_{(s,k),l}h_{l,r} \tag{4}$$

under the constraints

$$\begin{cases} \sum_k w_{(s,k),l} = 1, & \forall s,\forall l \\ \sum_l h_{l,r} = 1, & \forall l \end{cases} \tag{5}$$

This maximum likelihood formulation is equivalent to minimizing the extended Kullback-Leibler divergence between a matrix $\mathbf{V}$ and $\mathbf{WH}$ with $v_{(s,k),r} = \sum_t \gamma_{r;sk}(t)$. The update rules for $\mathbf{W}$ and $\mathbf{H}$ are the same as in [8], except for an extra state-wise L_1 normalization of $\mathbf{W}$ after each iteration.

The adapted Gaussian mixture weights for the evaluation speaker e are written as

$$\lambda_{e;sk} = \sum_{l=1}^{L} w_{(s,k),l}h_{e;l} \tag{6}$$

The latent vector coefficients h_e of the evaluation speaker e are estimated to maximize the enrollment data of that speaker, which results in the following iterative update rule:

$$h_{e;l} \leftarrow \sum_{s,k,t}\frac{\gamma_{e;sk}(t)w_{(s,k),l}}{\sum_{i=1}^{L} w_{(s,k),i}h_{e;i}}h_{e;l} \tag{7}$$

with $h_{e;l}$ L_1 normalized.

Analogous to eigenvoices, the latent weight vectors $\mathbf{W}$ encode relations. Whereas eigenvoices encode relations between Gaussian means, the latent weight vectors model the relations among the Gaussian mixture weights observed over the different speakers in the training data. Hence, during adaptation, unobserved Gaussian mixture weights can be inferred from the observed ones. The number of degrees of freedom equals the number of latent vectors minus one (the Gaussian mixture weights are probabilities which sum up to one), which is small. Therefore, the NMF weight-based speaker adaptation can adapt rapidly.

4. GAUSSIAN MEAN AND WEIGHT ADAPTATION

To improve the performance of the speech recognition system, we combine the eigenvoice and NMF weight-based rapid speaker adaptation techniques. The two techniques can be expected to combine very well for several reasons. i) Both eigenvoice speaker adaptation and NMF weight-based speaker adaptation maximize the likelihood of the enrollment data. ii) NMF weight-based speaker adaptation provides an elegant solution to the problem that eigenvoice speaker adaptation technique cannot be readily used to adapt the Gaussian mixture weights. iii) For both techniques, the acoustic models of the evaluation speaker are expressed as linear combinations of a set of latent vectors, which are estimated from the speaker dependent

acoustic models of the training speakers. iiii) Both techniques are rapid speaker adaptations with a limited number of parameters to be estimated based on the enrollment data.

Speaker adaptive training (SAT) [9] is used together with this adaptation algorithm. SAT improves the performance of the speech recognition system by reducing the inter-speaker variability and generating an acoustic model which more accurately represents the phonetic variations in the training data. In [7], we also showed that SAT helps in exposing more relevant relations between Gaussian mixture weights. The SAT model parameters λ_{SAT}, μ_{SAT}, Σ_{SAT}, and a common cMLLR transformation matrix $\mathbf{M}_{\text{SAT}}$ estimated on all training speakers jointly, are estimated using the maximum likelihood criterion. With only a few seconds of enrollment data, the speaker specific cMLLR estimate $\mathbf{M}_e$ was found to be unreliable and hence $\mathbf{M}_{\text{SAT}}$ is applied as the transformation matrix during adaptation and evaluation. Algorithm 1 lists the different steps involved in deriving the eigenvoices and latent weight vectors.

Algorithm 1 Deriving the latent vectors for adaptation.

Step 1 : Initialize the SD mean vectors $\{\mu_r\}_{r=1}^{R}$ and Gaussian mixture weights $\{\lambda_r\}_{r=1}^{R}$ of the training speaker:
$\mu_r = \mu_{\text{SAT}}$; $\lambda_r = \lambda_{\text{SAT}}$ with $r = 1 \cdots R$

Step 2 : Estimate the R SD mean vectors $\{\mu_r\}_{r=1}^{R}$ of the training speakers, using $\mathbf{M}_{\text{SAT}}$ as the transformation matrix.

Step 3 : Reestimate the SD mean vectors using MAP.

$$\mu_{r;sk}^{(\text{MAP})} = \frac{\gamma_{r;sk}}{\gamma_{r;sk}+\Gamma}\mu_{r;sk} + \frac{\Gamma}{\gamma_{r;sk}+\Gamma}\mu_{\text{SAT};sk} \tag{8}$$

where $\gamma_{r;sk}$ is the cumulative Gaussian posterior probability; Γ is the parameter to control the relative weight of the prior μ_{SAT}.

Step 4 : Estimate the eigenvoices $\{\phi_j\}_{j=1}^{J}$ by applying PCA to the correlation matrix of the Gaussian means $\{\mu_r^{(\text{MAP})}\}_{r=1}^{R}$.

Step 5 : Estimate the eigenvoice weighting coefficients $\{\rho_j\}_{j=2}^{J}$ of the training speakers by maximizing the likelihood of the data. The coefficients can be computed from equation (2). This step is iterated until the coefficients $\{\rho_j\}_{j=2}^{J}$ converge.

Step 6 : Estimate the SD weight vectors $\{\lambda_r\}_{r=1}^{R}$ of the training speakers using the adapted mean vectors estimated in *Step 5*.

Step 7 : Formulate the NMF $\mathbf{V}$ matrix. The rth SD weight vector λ_r forms the rth column of $\mathbf{V}$.

Step 8 : Estimate the latent vectors $\mathbf{W}$ using NMF by maximizing the likelihood of the training data; see [7] for more details.

In the eigenvoice technique, only the weights ρ_j are estimated in the ML-sense (step 5). The eigenvoices ϕ_j themselves are created using PCA (step 4). As a result, the eigenvoice model parameters ϕ_j and ρ_j are not jointly ML. Given this limitation, iterating steps 2 to 8 to jointly optimize all parameters, i.e. ϕ_j, ρ_j for the eigenvoices and $\mathbf{W}, \mathbf{H}$ for the NMF weight adaptation, would only make sense with ML-based eigenvoices as presented in [10]. Given that both the eigenvoices and the latent weight vectors are designed to approximate the original SD parameters as good as possible, we do not expect substantial changes in the Gaussian means and weights in such an iterative scheme.

During evaluation, the eigenvoice weighting coefficients $\{\rho_j\}_{j=2}^{J}$ and the NMF latent vector coefficients h_e of the evaluation speakers are estimated by maximizing the likelihood of the enrollment data, given that the adapted means and weights are expressed as linear combinations of latent vectors (equation (1) and equation (6)). Maximizing the likelihood is equivalent to maximizing the auxiliary function using Expectation-Maximization (EM).

$$Q(h_e, \rho_j) = Q(h_e) + Q(\rho_j) \tag{9}$$

with

$$\begin{aligned} Q(h_e) &= \textstyle\sum_{s,k,t} \gamma_{sk}(t) \log(\sum_l w_{(s,k),l} h_{e;l}) \\ Q(\rho_j) &= -\tfrac{1}{2} \textstyle\sum_{s,k,t} \gamma_{sk}(t) \big\{ n\log(2\pi) + \log|\Sigma_{\text{SAT};sk}| + \\ &\quad (\mathbf{o}_t - \textstyle\sum_{j=1}^{J} \phi_j \rho_j)^T \Sigma_{\text{SAT};sk}^{-1} (\mathbf{o}_t - \textstyle\sum_{j=1}^{J} \phi_j \rho_j) \big\} \end{aligned} \tag{10}$$

with observations $\mathbf{O} = \mathbf{o}_1 \cdots \mathbf{o}_T$ and n the feature dimension. The optimum can be found by iterating equation (2) (with $\mu_{\text{SI};sk} = \mu_{\text{SAT};sk}$ and $\Sigma_{\text{SI};sk} = \Sigma_{\text{SAT};sk}$) and equation (6) until the coefficients converge.

5. EXPERIMENTAL RESULTS

5.1. Recognition system

The adaptation algorithms are evaluated on the Wall Street Journal (WSJ) database. Training is done on the SI-284 data from both the WSJ0 and WSJ1 database comprising 81 hours speech from 284 speakers. The acoustic model uses a shared pool of 32754 Gaussians to model the observations in 5967 cross-word context-dependent tied triphone states. On average, 94 Gaussian probability densities are used to describe the emission probabilities per state. We use a 3-state left-to-right topology to describe all the acoustic units — context-dependent variants of one of the 42 phones or silence. Mean normalization and VTLN are included in the preprocessing of the recognition system. The acoustic features consist of 22 MEL spectra, augmented with their first and second order time derivatives, which generates 66 dimensional feature vectors. By means of a discriminative linear transformation and decorrelation, these features are then mapped to a lower 39 dimensional space. Notice that unlike in other work, we adapt Gaussians that are tied over states.

For both development and evaluation, the WSJ 5k closed and 20k open vocabulary non-verbalized punctuation Nov92 and Nov93 tasks are combined. The development data are used to tune the system parameters such as the pruning thresholds and the weight ratio between the language model and the acoustic model. By combining all the evaluation data, we obtain one large evaluation set containing 101 minutes of speech (18298 words).

The adaptation in the experiments is supervised. Given that the same speakers are shared in each pair of 5k and 20k sub-tasks, we choose the enrollment data outside the current sub-task, i.e. draw enrollment data from the 5k sub-task to evaluate on the data of the same speaker in the 20k sub-task and vice versa. By this way, we can investigate the proposed algorithm with variable amounts of enrollment data.

The recognition system uses a 75k lexicon and a standard trigram language model. The out-of-vocabulary ratio on the development and evaluation set are 0.08% and 0.12% respectively. In our experiments, Γ (equation (8)) is set to 3.

5.2. Results and discussions

Table 1 shows the word error rate (WER) on the evaluation data for different speaker adaptation algorithms. In a preliminary experiment we verified that the SAT acoustic model performed as good as the SI model (which gives a 6.42% WER). In fact, we observed better results in most tests when using the SAT model instead of the SI model. For the NMF decomposition, the cumulative posteriors corresponding to the 3-state silence model are discarded from $\mathbf{V}$. Hence, the Gaussian mixture weights of the silence states retain their corresponding SAT mixture weight values.

# latent vectors		WER for different amounts of enrollment data (seconds)				
J	L	0	3	6	8	100+
0	0	6.37	/	/	/	/
0	2	/	6.16	6.15	6.14	6.13
2	0	/	6.01	6.02	5.97	6.00
2	2	/	6.03	6.00	5.93	5.97
0	10	/	6.06	6.08	6.09	6.01
10	0	/	5.95	5.99	5.90	5.97
10	10	/	5.87	5.89	5.76	5.81
0	40	/	6.01	6.06	6.08	6.03
40	0	/	6.15	6.04	6.00	6.04
40	40	/	6.01	5.88	5.78	5.69

Table 1. Evaluation data WER (%) obtained with different adaptation algorithm. *100+*: the amount of enrollment data per speaker is around 240 seconds for Nov92 and 100 seconds for Nov93 .

Overall we see that both eigenvoice and weight adaptations require 10 or less latent vectors to express the speaker variability. Here we should remark that the preprocessing contains VTLN and thus removes one of the major factors in inter-speaker variability.

The fact that with only 2 latent vectors, eigenvoices performs better than NMF could be related to the lower dimensionality of the latent vectors for NMF (weights) compared to those for eigenvoices (all Gaussian means). The lower dimensionality may result in somewhat less expressive latent vectors. Once sufficient latent vectors are used, the two methods perform almost identical.

The larger dimensionality of the eigenvoices may also make the eigenvoices more susceptible to overfitting when only limited amounts of enrollment data are available. This may explain the increase in WER when using 40 instead of 10 eigenvoices with only 3 seconds of enrollment data for the eigenvoice system, whereas the weight adaptation (with lower dimensional latent vectors) shows no signs of overfitting.

Combining the two methods with only 2 latent vectors shows no improvement over the eigenvoice approach. The most likely reason for this is that both approaches use their single degree of freedom to model the same inter-speaker variability. However, with more degrees of freedom, combining the two adaptation schemes shows consistent improvements over both the eigenvoice and the weight adaptation scheme on itself. For example, with 10 degrees of freedom and 3 seconds of enrollment data, the performance of the recognition system is improved by 1.3% relatively over the best single adaptation method, resulting in a 7.9% improvement compared to the SAT baseline. When more degrees of freedom (40) and more enrollment data (100+) are available, this improvement increases to 5.8% compared to the eigenvoice approach and 10.7% compared to the SAT baseline. Furthermore, the combination with less degrees of freedom ($J = 10, L = 10$) gives better results than either the eigenvoice or the weight adaptation with more degrees of freedom ($J = 40, L = 0$ and $J = 0, L = 40$ respectively).

These results show that NMF-based Gaussian mixture weight adaptation and eigenvoice-based mean adaptation are compatible with each other. NMF weight-based adaptation is complementary to eigenvoice speaker adaptation. By applying the combination of these two adaptation techniques, more speaker related information is available. We observe that the eigenvoice adaptation keeps the Gaussians that are unneeded for a certain training speaker close to their SI positions. This may create a source for undesirable overlap of state densities. The NMF weight adaptation can suppress these Gaussians by assigning a relative small value or zero to the corresponding Gaussian mixture weights.

6. CONCLUSIONS

This paper described a novel model space fast speaker adaptation algorithm which adjusts both the Gaussian means and the Gaussian mixture weights. The adapted Gaussian means are expressed as a linear combination of eigenvoices. Similarly, the Gaussian mixture weights are expressed as a weighted sum over a set of latent vectors. By exploiting the pre-learned Gaussian mean and Gaussian mixture weight relations, the statistics for both observed and unobserved acoustic model parameters are updated, resulting in good generalization. By expressing the model parameters in function of a small set of eigenvoices or latent weight vectors, the degrees of freedom are reduced, resulting in fast adaptation.

The two ML-based techniques are compatible with each other. With sufficient degrees of freedom, the combination of these two techniques outperforms the eigenvoice speaker adaptation technique alone consistently for both large or small amounts of enrollment data. The combination with less degrees of freedom yields better performance than the eigenvoice or NMF weight adaptation alone with more degrees of freedom.

7. REFERENCES

[1] C. Tuerk and T. Robinson, "A new frequency shift function for reducing inter-speaker variance," in *Proc. Eurospeech*, 1993, pp. 351–354.

[2] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. on SAP*, vol. 2, pp. 291–298, 1994.

[3] G. Zavaliagkos and R. Schwartz, "Maximum a posteriori adaptation for large scale HMM recognizers," in *Proc. ICASSP*, 1996, vol. 2, pp. 725–728.

[4] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comp. Speech and Lang.*, vol. 12, pp. 75–98, 1998.

[5] D. Povey and K. Yao, "A basis method for robust estimation of constrained MLLR," in *Proc. ICASSP*, 2011, pp. 4460–4463.

[6] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on SAP*, vol. 8, no. 6, pp. 695–707, 2000.

[7] X. Zhang, K. Demuynck, and H. Van hamme, "Rapid speaker adaptation with speaker adaptive training and non-negative matrix factorization," in *Proc. ICASSP*, May 2011, pp. 4456–4459.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, pp. 556–562, 2001.

[9] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. International Conference in Spoken Language Processing*, 1996, pp. 1137–1140.

[10] P. Nguyen, C. Wellekens, and J.-C. Junqua, "Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments," in *Eurospeech*, 1999, pp. 2519–2522.