RAPID SPEAKER ADAPTATION WITH SPEAKER ADAPTIVE TRAINING AND NON-NEGATIVE MATRIX FACTORIZATION

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

In this paper, we describe a novel speaker adaptation algorithm based on Gaussian mixture weight adaptation. A small number of latent speaker vectors are estimated with non-negative matrix factorization (NMF). These base vectors encode the correlations between Gaussian activations as learned from the train data. Expressing the speaker dependent Gaussian mixture weights as a linear combination of a small number of base vectors, reduces the number of parameters that must be estimated from the enrollment data. In order to learn meaningful correlations between Gaussian activations from the train data, the NMF-based weight adaptation was combined with vocal tract length normalization (VTLN) and feature-space maximum likelihood linear regression (fMLLR) based speaker adaptive training based. Evaluation on the 5k closed and 20k open vocabulary Wall Street Journal tasks shows a 4% relative word error rate reduction over the speaker independent recognition system which already incorporates VTLN. The proposed fast adaptation algorithm, using a single enrollment sentence only, results in similar performance as fMLLR adapting on 40 enrollment sentences.

Index Terms— Speaker adaptation, non-negative matrix factorization, speaker adaptive training, maximum likelihood linear regression, weight adaptation

1. INTRODUCTION

Given a sufficient amount of *speaker dependent* (SD) training data, speaker dependent speech recognizers generally perform better than their *speaker independent* (SI) counter parts [1]. However, for most applications, only limited amounts of speaker dependent data are available, insufficient to make a true speaker dependent system. Examples thereof are speech based automatic vending machines or automatic telephone services. Under these circumstances, *speaker adapted* models form an appealing solution. Speaker adapted models transform the SI acoustic model so that, given some limited amounts of example data from that speaker, the adjusted acoustic model better describes the target speaker's speech.

In the last decades, speaker adaptation techniques have focused on feature-space and model-space transformations of the Gaussian means (and variances). Little attention has been given to the Gaussian mixture weights. In this paper, we focus on fast Gaussian mixture weights based model-space speaker adaptation. Speaker adaptation techniques are typically characterized on the following characteristics: (i) generalization, i.e. can the model parameters (context-dependent phone distributions) for which no or little enrollment data was observed (henceforth "unseen parameters") be derived from those model parameters that were observed, (ii) susceptibility to overfitting when only small amounts of enrollment data are available, and (iii) convergence to the true SD model with infinite amounts of data.

Maximum a posteriori (MAP) adaptation maximizes the posterior probability of the model parameters given the adaptation data, with the SI acoustic model parameters used as priors [2]. MAP estimations converge to the maximum likelihood (ML) estimations if infinite amounts of enrollment data are provided. The priors counteract overfitting. A disadvantage of the MAP algorithm is that it does not generalize: only the observed parameters are updated, the unseen parameters retain their SI model parameter values.

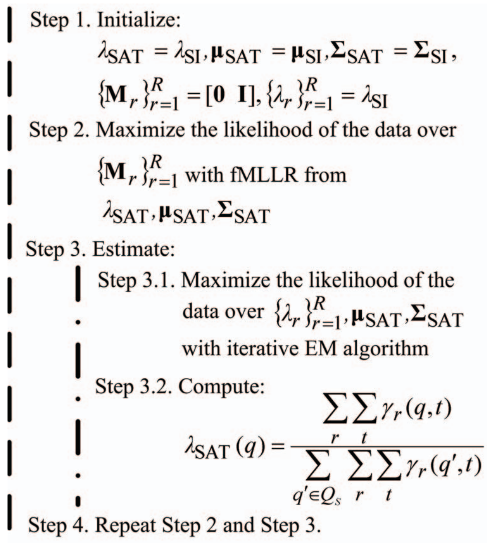
Eigenvoice speaker adaptation [3] expresses the Gaussian means as a linear combination of eigenvoices of Gaussian means. The eigenvoices are learned by means of eigenvalue decomposition of the SD Gaussian means for the train speakers. By exploiting the correlations between Gaussian means, as encoded in the eigenvoices, this method can, based on very small amounts of enrollment data, infer the Gaussian means reliably for both seen and unseen distributions. Combination with MAP allows convergence to the true SD model with infinite amounts of data.

Unconstrained and constrained (feature-space) maximum likelihood linear regression (UMLLR/fMLLR) [4] maximize the likelihood of the enrollment data when allowing linear transformations of the Gaussian means and variances. Generalization is largely dependent on whether a linear transformation is a good model to characterize inter-speaker differences. When convergence to the true SD model is required, the number of transformations must be increased when more data becomes available. To avoid overfitting, the number of free parameters in the linear transformations must be limited if insufficient amounts of enrollment data are available, for example by using eigenspaces [5].

In this paper, we look at adaptation of the Gaussian mixture weights instead of the more common Gaussian mean and variance adaptation. Similar to eigenvoice speaker adaptation, a set of base vectors expressing correlations learned from the speakers in the train database is used to provide generalization of unseen parameter distributions from seen distributions. The base vectors are obtained with non-negative matrix factorization (NMF) [6] of Gaussian posteriors as recorded on the train speakers, similar to what was done in [7]. Next to providing a good generalization, NMF weight adaptation also requires very small amounts of enrollment data. Estimating the cumulative Gaussian posteriors (the input for the NMF adaptation) is even less complex than estimating the Gaussian specific first order feature moments required by the eigenvoice method, and hence can be done with similar amounts of enrollment data. Given the tendency of modern hidden Markov model (HMM) systems to use large Gaussian mixtures to model the observation probability density distribu-

[*] This author is also affiliated with IBBT (Interdisciplinary Institute for Broadband Technology).

978-1-4577-0539-7/11/$26.00 ©2011 IEEE



```
Step 1. Initialize:
        $\lambda_{SAT} = \lambda_{SI}, \boldsymbol{\mu}_{SAT} = \boldsymbol{\mu}_{SI}, \boldsymbol{\Sigma}_{SAT} = \boldsymbol{\Sigma}_{SI}$,
        $\{\mathbf{M}_r\}_{r=1}^{R} = [\mathbf{0}\ \mathbf{I}], \{\lambda_r\}_{r=1}^{R} = \lambda_{SI}$
Step 2. Maximize the likelihood of the data over
        $\{\mathbf{M}_r\}_{r=1}^{R}$ with fMLLR from
        $\lambda_{SAT}, \boldsymbol{\mu}_{SAT}, \boldsymbol{\Sigma}_{SAT}$
Step 3. Estimate:
    Step 3.1. Maximize the likelihood of the
              data over $\{\lambda_r\}_{r=1}^{R}, \boldsymbol{\mu}_{SAT}, \boldsymbol{\Sigma}_{SAT}$
              with iterative EM algorithm
    Step 3.2. Compute:
              $\lambda_{SAT}(q) = \frac{\sum_r \sum_t \gamma_r(q,t)}{\sum_{q' \in Q_s} \sum_r \sum_t \gamma_r(q',t)}$
Step 4. Repeat Step 2 and Step 3.
```

Fig. 1. Speaker adaptive training in the proposed speaker adaptation algorithm. The dashed lines show the outer loop of the SAT. The dash dotted lines show the inner loop of the SAT process.

tions (pdfs), weight adaptation is also surprisingly flexible: complex non-linear transformations can be obtained by just changing weights. We also investigate the effect of NMF-based weight adaptation in combination with state-of-the-art speaker normalization and adaptation techniques such as vocal tract length normalization (VTLN), fMLLR, and speaker adaptive training (SAT) [8]. In fact, SAT with good speaker normalization and adaptation schemes proved to be highly favorable since this resulted in more active Gaussians per speaker and hence allowed the NMF-decomposition to record more meaningful correlations between the observed Gaussian activations.

The remainder of this paper is organized as follows. In section 2, we introduce SAT based on speaker dependent fMLLR matrices, a single shared set of Gaussians, and speaker dependent Gaussian mixture weights. Section 3 recapitulates the NMF based speaker adaptation algorithm. We describe our recognition system and compare the recognition results with different speaker adaptation algorithms in section 4. In section 5, conclusions and possible future research topics are presented.

2. SPEAKER ADAPTIVE TRAINING

Speaker adaptive training improves the performance of speech recognition systems by reducing the inter-speaker variation and meanwhile more accurately representing the phonetic variation in the training data [8]. Figure 1 shows the SAT as used in combination with our NMF-based speaker adaptation algorithm. Let $\lambda, \mu, \mathbf{\Sigma}, \mathbf{M}$ represent the Gaussian mixture weights, mean vector, variance matrix, and fMLLR extended linear transformation matrix of the acoustic model respectively. Subscripts SI, SAT, and r represent the SI acoustic model, SAT estimated acoustic model, and the rth training speaker (reference speaker). R is total number of reference speakers, t the time index, s the state index, q the Gaussian component index, Q_s the set of component indices for a given state s, and $\gamma_r(q,t)$ the posterior probability of the observation for Gaussian q at time t using the SAT model of speaker r. The SAT model parameters $\{\lambda_r\}_{r=1}^{R}$, λ_{SAT}, μ_{SAT}, $\mathbf{\Sigma}_{\text{SAT}}$, and $\{\mathbf{M}_r\}_{r=1}^{R}$ are trained in a nested loop using maximum likelihood re-estimation. The outer loop optimizes the feature-space transformation matrices. In *Step 2*, fMLLR [4] is used to estimate the SD matrices $\mathbf{M}_r$ based on the data of the individual speakers and based on the current estimate of the SAT model, i.e μ_{SAT}, $\mathbf{\Sigma}_{\text{SAT}}$ and λ_{SAT}. For reasons we will explain later on, we also estimate a common fMLLR transformation matrix $\mathbf{M}_{\text{SAT}}$ on all speakers jointly using μ_{SAT}, $\mathbf{\Sigma}_{\text{SAT}}$ and λ_{SAT}, the latter being the common mixture weights as formed in *Step 3.2*. Transforming the observation vectors $\hat{\mathbf{o}}(t) = \mathbf{A}\mathbf{o}(t) + \mathbf{b}$ with $\mathbf{M} = [\mathbf{b}\ \mathbf{A}]$ and $\mathbf{A}$ a full transformation matrix, allows the inner loop to use the standard expectation-maximization (EM) algorithm to update the Gaussian distributions and mixture weights (*Steps 3.1+3.2*). The SD Gaussian mixture weights $\{\lambda_r\}_{r=1}^{R}$ are used to perform NMF based mixture weight speaker adaptation as described in the next section.

3. GAUSSIAN MIXTURE WEIGHT ADAPTATION USING NON-NEGATIVE MATRIX FACTORIZATION

In our speaker adaptation algorithm, the Gaussian mixture weights of the SI/SAT model are adapted toward the evaluation (target) speaker. The weights are adjusted through a NMF based speaker adaptation algorithm described in detail in [7]. Here we recapitulate the fundamentals of the adaptation algorithm. NMF [6] approximates a non-negative matrix $\mathbf{V}$ as the product of two non-negative matrices: a basis vector matrix $\mathbf{W}$ and a coefficient matrix $\mathbf{H}$. For the NMF speaker adaptation, the matrices $\mathbf{W}$ and $\mathbf{H}$ are chosen to maximize the data likelihood

$$Q(W,H) = \sum_r \sum_q \sum_t \gamma_r(q,t) \log(\lambda_r(q)), \tag{1}$$

with

$$\lambda_r(q) = \frac{\sum_l W_{q,l} H_{l,r}}{\sum_l H_{l,r}}, \tag{2}$$

under the constraints

$$\sum_{q \in Q_s} W_{q,l} = 1, \qquad \text{for all states } s, \text{ and base vectors } l \tag{3}$$

Except for the state wise normalization of W, maximizing eqn. (1) given the constraint (3) is equivalent to minimizing the extended Kullback-Leibler divergence between a matrix $\mathbf{V}$ and $\mathbf{WH}$:

$$D(V||WH) = \sum_{q,r} (V_{q,r} \log \frac{V_{q,r}}{(WH)_{q,r}} - V_{q,r} + (WH)_{q,r}) \tag{4}$$

with $\mathbf{V}_{q,r} = \sum_t \gamma(q,t)$. Hence, we found the same update rules for W and H as in [6], except for an extra state-wise L_1 normalization of W after each iteration.

Given the latent speaker matrix $\mathbf{W}$, the latent coefficients h_e of a target (enrollment) speaker e are estimated iteratively based on the EM algorithm (Baum-Welch) given the enrollment data of that speaker. The ith iteration of the latent speaker coefficient is given by equation (5) (see [7] for more details):

$$h_e^{i+1}(l) = \sum_q \sum_t \frac{\gamma_e(q,t) W(k,l)}{\sum_{j=1}^{L} W(q,j) h_e^i(j)} \times h_e^i(l) \tag{5}$$

where $\gamma_e(q,t)$ are obtained by Viterbi alignment of the evaluation speaker's data using λ_{SAT}, μ_{SAT}, $\mathbf{\Sigma}_{\text{SAT}}$, $\mathbf{M}_e$/$\mathbf{M}_{\text{SAT}}$. If sufficient enrollment data are available, $\mathbf{M}_e$ is estimated and used. Otherwise, $\mathbf{M}_{\text{SAT}}$ is used. In case of unsupervised adaptation, $\gamma_e(q,t)$ are obtained from alignment on the best hypothesis (see section 4.1).

The product of the latent speaker matrix $\mathbf{W}$ and the estimated coefficient vector $\mathbf{h}$ is the estimated SA Gaussian mixture weight vector for the target speaker.

4. EXPERIMENTAL RESULTS

4.1. Recognition system

The Wall Street Journal (WSJ) corpus is used for training, developing and testing the NMF speaker adaptation. Training is done on the SI-284 data from WSJ0+1 comprising 81 hours from 284 speakers. The baseline speech recognizer used in our experiments is a semi-tied Gaussian mixture HMM system. The system uses a shared pool of 32754 Gaussians to model the observations in 5967 cross-word context-dependent tied triphone states, using 94 Gaussian pdfs on average per state. All acoustic units –context-dependent variants of one of the 42 phones or silence– have a 3-state left-to-right topology. The acoustic features consist of Mel Spectra with mean normalization and VTLN [9], augmented with their first and second order time derivatives. These features are then mapped to a 39 dimensional space by means of a discriminative linear transformation [10] and decorrelated [10].

Both the 5k closed and 20k open vocabulary corpora, each with 8 speakers, in the Nov92 evaluation data are used for adaptation and evaluation. Identical to [7], we reserved the first 10 sentences of each speaker for enrollment (of which either one or all ten are actually used). We evaluate on the remaining (±30) sentences, henceforth called "test data". For tests with more (±40) enrollment sentences, we evaluate on the same set, but extract the enrollment data from the evaluation set with the other vocabulary size, i.e. enroll on the 5k vocabulary, test on the 20k vocabulary and vice versa.

The proposed fast speaker adaptation algorithm has been tested both supervised (the enrollment data transcription is known) and unsupervised (the enrollment data transcription is unknown and must be estimated). For unsupervised adaptation, a two-pass scheme is required. First, the recognizer processes the enrollment data, generating both single best word sequences and word lattices. Word posteriors are derived from the word lattices. Based on the word posteriors, the 30% least likely words from the single best word sequences are discarded. From there on, the recognition proceeds as with supervised adaptation: Viterbi alignment is used to find the best state alignment path, which is then used to estimate the Gaussian mixture weight posterior probabilities and/or the fMLLR statistics needed for the adaptation. Note that when combining SAT and NMF adaptation with only a single adaptation sentence, the fMLLR estimate $\mathbf{M}_e$ was found to be unreliable and hence was replaced with $\mathbf{M}_{\text{SAT}}$ during adaptation and evaluation.

4.2. Results

Table 1 gives the word error rate (WER) in % for the different speaker adaptation algorithms. The 5k and 20k test corpus are both evaluated with their respective bigram (2g) and trigram (3g) language models, resulting in four configuration: *5k-2g*, *5k-3g*, *20k-2g*, and *20k-3g*. For the NMF decomposition, the number of latent speaker vectors is set to 10 and the model parameters corresponding to the 3-state silence model are excluded, i.e. they retain their SI/SAT model parameter values.

The standard deviation std on the WER for the four conditions, calculated by eqn. (6) with N_{word} the number of words in the reference transcription, equals to: ±0.30% (*5k-2g*), ±0.44% (*20k-2g*), ±0.23% (*5k-3g*), and ±0.40% (*20k-3g*). Given this relatively large standard deviations, one can best look at results averaged over all four conditions when comparing methods.

$$std = \sqrt{\frac{\text{WER} * (1 - \text{WER})}{N_{\text{word}}}} \tag{6}$$

HMM	adaptation method	5k-2g	20k-2g	5k-3g	20k-3g
SI	/	3.74	8.97	2.21	7.35
SI	NMF[1]	3.72	8.75	2.28	7.30
SI	NMF[10]	3.79	8.61	2.31	7.32
SI	fMLLR[1]	3.67	9.23	2.48	7.61
SI	fMLLR[40]	3.81	8.66	2.53	7.18
SAT	/	3.76	9.11	2.43	7.13
SAT	fMLLR[40]	3.62	8.32	2.31	7.06
SAT	NMF[40]+fMLLR[40]	3.38	8.47	2.21	6.96
SAT	NMF[1]	3.62	8.66	2.23	6.85
SAT	NMF[1]; unsupervised	3.64	8.51	2.23	6.85
SI	NMF[1] (SAT W)	3.69	8.62	2.31	7.17
SI	NMF[10] (SAT W)	3.67	8.67	2.28	7.09

Table 1. WER(%) obtained with different acoustic models and speaker adaptation algorithms. The number between square brackets is the number of enrollment sentences used for the adaptation.

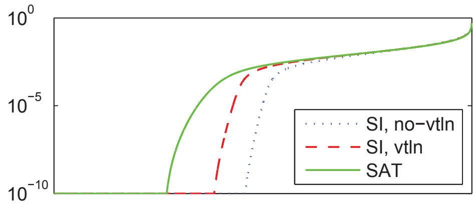


Fig. 2. Distribution of the values in the NMF latent speaker vectors.

As can be seen from table 1, NMF weight adaptation on the baseline SI acoustic model (*SI+NMF[1]* and *SI+NMF[10]*) does not provide a significant WER improvement over the SI baseline. This contrasts with the results reported in [7] which show a 5% to 15% relative improvement in this situation. However, the setup in [7] does not include VTLN in the front-end preprocessing. Consequently the NMF based speaker adaptation in [7] predominantly adapts the acoustic model to the speaker gender. In other words, in [7], the NMF based speaker adaptation process mainly plays the role of VTLN. On the other hand, in our setup, any observed improvement is an improvement in addition to VTLN.

Switching to the SAT+NMF setups shows that significant improvements can be obtained with the NMF-based weight adaptation scheme. The key assumption underlying the NMF adaptation technique is that meaningful correlations between Gaussian activations can be learned from the train data. With SAT, each speaker-dependent HMM state will simply have more active Gaussians, hence providing richer information to the NMF decomposition. This results in better latent base vectors W. This is illustrated in figure 2. Given the non-negative nature of the values in W, the only way correlations between Gaussians activations can be encoded is by having a positive value for the respective Gaussians. Hence, having more non-zero weights is a strong indicator that more (meaningful) correlations are encoded in W. As can be seen from the figure, SAT clearly improves upon the SI model with vtln in that aspect. As will be show later on, without vtln, the recorded correlations are even reduced to those related to the speaker's gender. A second indication that the driving factor for the improvements observed when going from a SI+NMF to a SAT+NMF setup, is the quality of the latent base vectors W is given by the good results obtained with the last two setups listed in table 1. For these experiments, we used the latent base vectors W learned from the SAT system in combination with the SI system.

	HMM providing the latent vectors			
meta-data	priors	SI, no vtln	SI, vtln	SAT
gender	50%	99%	94%	95%
age	38%	41%	48%	51%
region	57%	57%	54%	57%

Table 2. Classification accuracy for different speaker characteristics.

Table 1 also lists several results with fMLLR speaker adaptation. Comparing the different setups shows that: (i) without SAT, fMLLR does not yield additional improvements over VTLN, (ii) with SAT and with sufficient amounts of enrollment data, fMLLR lowers the WER on average by 4%, (iii) fMLLR and NMF-adaptation show similar (4% relative) improvements, and (iii) combining fMLLR and NMF-adaptation results in a small additional improvement.

When the amount of fMLLR adaptation data is limited to 1 sentence (*SI+fMLLR[1]*), it is no longer possible to reliably estimate the speaker dependent transformation matrix $\mathbf{M}_e$. We hence replace $\mathbf{M}_e$ with $\mathbf{M}_{\text{SAT}}$ and find that the NMF speaker adaptation benefits are maintained (*SAT+NMF[1]*), even in the unsupervised scenario.

When NMF is applied to the SI or SAT baseline system, it gives similar performance with a single enrollment sentence or 10 enrollment sentences. This shows that (i) NMF is not susceptibility to overfitting when only small amounts of enrollment data are available, and (ii) NMF generalizes well and hence can work with small amounts of adaptation data. Even unsupervised adaptation can be done with a single enrollment sentence. fMLLR adaptation (with a full transformation matrix) on the other, clearly requires more that one adaptation sentence (*SI+fMLLR[1]* versus *SI+fMLLR[40]*).

We also investigated the relations among the latent speaker vectors and some observable speaker characteristics. Identical to [7], the analysis was limited to the 200 WSJ1 training speakers for which speaker meta-data is available. The following speaker characteristics (with corresponding classes and counts) were considered: *gender* (male: 100, female: 100), *age* (< 25: 26, < 35: 75, < 45: 55, < 55: 26, ≥ 55: 16), and *region* where the speaker went to primary school (West: 110, the Midlands: 27, Southern: 19, New England: 16, the Inland North: 9, New York City: 6, North Central: 5). The classification was done with a simple linear classifier trained on the H matrix (the latent speaker coefficients) in a leaving-one-out scheme. Table 2 shows the results. Classification based on the prior class distributions and based on the latent speaker coefficients for a SI model without VTLN are added for reference. Without VTLN, *gender* is the only discernible speaker characteristic. Adding VTLN and SAT, allows the NMF-decomposition to encode additional speaker characteristics such as *age* in the base vectors. However, *gender* remains the most prominent feature, indicating that neither VTLN nor fMLLR can completely compensate for the gender differences. The investigation whether the latent speaker vectors also reflect speaker *accent*, measured by means of the *region* where the speaker went to primary school, was inconclusive. The lack of any observed correlation could be caused by several factors: (i) *region* may not predict *accent* since a great amount of the speakers moved among different regions, (ii) other characteristics such as ethnicity may be more related to accent, (iii) we work on read speech so most speakers adopt a relative standard English accent, and (iv) the latent vectors only have 10 degrees of freedom and already encode *gender* and *age* information; adding *accent* information may require more base vectors.

5. CONCLUSIONS AND FUTURE RESEARCH

This paper describes a novel fast speaker adaptation algorithm where SAT and NMF based speaker adaptation techniques are combined. It has been shown that these two maximum likelihood based techniques are compatible. By combining both techniques, the WER of the speech recognition system decreases on average 4% compared to the speaker independent baseline system. Unsupervised adaptation results in similar WER results as the supervised adaptation. The proposed Gaussian mixture weight adaptation algorithm requires little adaptation data (only one enrollment sentence).

A key characteristic of the proposed algorithm is that both the seen and unseen acoustic model parameters are updated by expressing the target speaker Gaussian mixture weights as a linear combination of latent speaker Gaussian mixture weight vectors. The performance of the weight adaptation is directly dependent on the quality of the correlations encoded in the latent vectors. We showed that SAT improves the quality of the latent vectors. Future research will focus on additional methods to improve the quality of the latent vectors. We also intend to apply hierarchical weight decomposition as to adjust the degrees of freedom in the NMF-adaptation to the amount of available adaptation data. Switching to a small number of base vectors (degrees of freedom) avoids the overfitting problem when little enrollment data is available. Increasing the number of base vectors with large amounts of enrollment data allows the system to get closer to the true speaker dependent model.

6. REFERENCES

[1] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall PTR, 2001.

[2] Phil. C. Woodland, “Speaker adaptation for continuous density HMMs: A review,” *ITRW on adaptation methods for speech recognition*, pp. 11–19, Aug. 2001.

[3] Roland Kuhn, Jean-Claude Junqua, Patrick Nguyen, and Nancy Niedzielski, “Rapid speaker adaptation in eigenvoice space,” *IEEE Trans. on SAP*, vol. 8, no. 6, pp. 695–707, 2000.

[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] Kuan-ting Chen, Wen-wei Liau, Hsin-min Wang, and Lin-shan Lee, “Fast speaker adaptation using eigenspace-based maximum likelihood linear regression,” in *Proc. ICSLP*, 2000, vol. 3, pp. 742–745.

[6] Daniel. D. Lee and H. Sebastian. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in Neural Information Processing Systems*, 2001, vol. 13, pp. 556–562.

[7] Jacques Duchateau, Tobias Leroy, Kris Demuynck, and Hugo Van hamme, “Fast speaker adaptation using non-negative matrix factorization,” in *Proc. ICASSP*, Apr. 2008, pp. 4269–4272.

[8] Tasos Anastasakos, John McDonough, and John Makhoul, “Speaker adaptive training: a maximum likelihood approach to speaker normalization,” in *Proc. ICASSP*, Apr. 1997, pp. 1043–1046.

[9] Jacques Duchateau, Mari Wigham, Kris Demuynck, and Hugo Van hamme, “A flexible recogniser architecture in a reading tutor for children,” in *ITRW on Speech Recognition and Intrinsic Variation*, May 2006, pp. 59–64.

[10] Kris Demuynck, *Extracting, Modelling and Combining Information in Speech Recognition*, Ph.D. thesis, Katholieke Universiteit Leuven, 2001.