SPEAKER ADAPTIVE TRAINING APPLIED TO CONTINUOUS MIXTURE DENSITY MODELING

Xavier Aubert, Eric Thelen

Philips GmbH Forschungslaboratorien Aachen, P.O. Box 50 01 45, D-52085 Aachen, Germany
E-mail: {aubert,thelen}@pfa.research.philips.com

ABSTRACT

Speaker Adaptive Training (SAT) has been investigated for mixture density estimation and applied to large vocabulary continuous speech recognition. SAT integrates MLLR adaptation in the HMM training and aims at reducing inter-speaker variability to get enhanced speaker-independent models. Starting from BBN's work on compact models, we derive a one-pass Viterbi formulation of SAT that performs joint estimation of MLLR-based transformations and density parameters. The computational complexity is analyzed and an approximation based on using inverse affine transformations is discussed.

Compared to applying MLLR on standard SI models, our experimental results achieve lower error rates as well as reduced decoding costs, for both supervised batch and unsupervised incremental adaptation. In the latter case, it is shown that the enrollment of a new speaker can be sped up by selecting among the transformations that were estimated from the training speakers, the one that best fits with the first test utterance.

1. INTRODUCTION

With the achievement of effective speaker adaptation techniques based on transformation methods, there has been a growing interest for applying similar techniques in training, to the speakers that are used for constructing speaker-independent (SI) acoustic models [2, 4, 5, 9, 10, 11].

As a matter of fact, the standard way of training SI models is to estimate the acoustic parameters from the "raw" speech of a large population of speakers and this leads to models with broad distributions requiring a large number of (typically) mixture densities. However, this does not necessarily provide the best acoustic models for a recognition system that includes a particular adaptation technique aimed at achieving speaker-dependent (SD) performances as fast as possible.

The motivation for normalizing the training speakers is to get rid of some of the inter-speaker variability so that the resulting SI models are more focussed on the phonetically relevant variation sources. Hence, they should exhibit sharper distributions with reduced variances and might be better suited for fast and accurate adaptation towards speaker-dependent models.

So far, two distinct ways to speaker normalization have been mainly considered. The first approach belongs to Vocal Tract Normalization (VTN) techniques and compensates for variations in vocal tract length by means of a (linear) warping of the frequency axis in the speech parameterization front-end (see a.o. [3] [10]).

The second avenue, which we are concerned with in this work, performs model adaptation using affine transformations of the density means. These transformations can be efficiently computed by the maximum likelihood linear regression (MLLR) algorithm [1] and have been shown to capture speaker specific characteristics reasonably well.

This leads to the so-called Speaker Adaptive Training (SAT) method recently introduced by Anastasakos et al. [5], where the parameters of both MLLR-based transformations and HMM mixture densities are *jointly* estimated in a unified maximum likelihood (ML) framework. As shown in [8], the resulting mixture distributions have markedly reduced variances compared to standard SI models. This raises the question of how these SAT models could be best used for decoding. So far, the most successful results have been obtained under supervised batch adaptation by combining the SAT models with MLLR-based transformations dedicated to the test speakers [5] [8] [10] [11]. This is not surprising as it allows to match the training conditions quite well via supervised MLLR estimation. However, for unsupervised incremental adaptation, a situation of clear practical interest, the SAT HMMs alone might not supply an ideal starting point. Indeed, in some cases at least, it appears that they perform worse than standard SI models. This suggests that we might take advantage of the transformations jointly estimated during training to get better initial models for a new speaker. We have addressed this possibility and our results indicate that the enrollment of a new speaker can be sped up by selecting among the transformations estimated from the training speakers, the one that best fits with the first test utterance.

In this paper, we first review the SAT equations, give a simple interpretation of the mean re-estimation formula and discuss an approximation based on using inverse affine transformations. Next, we address the computational issues and present our Viterbi SAT implementation. Development results are then presented under supervised adaptation to study the influence of regression class numbers and of approximating the SAT mean re-estimation. In the last section devoted to unsupervised adaptation, we present our first attempt upon using in decoding the MLLR transforms that were estimated during training, and give several experimental results.

2. JOINT ESTIMATION OF MLLR AND MIXTURE DENSITY PARAMETERS

In the sequel, we will associate index $r = 1, ..., R$ with the training speakers, index $g = 1, ..., G$ with the MLLR regression classes and $k = 1, ..., K$ with the mixture density components. The MLLR-based affine transformations are speaker-dependent and consist of full $(d \times d)$ matrices $\mathbf{A}_{r,g}$ and d-dimensional offset vectors $\mathbf{b}_{r,g}$. Each mixture component k is a Gaussian distribution $p(o_t|k) \sim \mathcal{N}(\mu_k, \Sigma_k)$, o_t being the observation vector at time t, μ_k and Σ_k being the mean vector and the (diagonal) covariance matrix.

For notational convenience, each speaker r is assumed to have produced a single utterance of length $T(r)$.

The basic idea underlying SAT is that the characteristics of each training speaker are carried by a set of linear transformations mapping the SI means on the speaker-specific acoustic domain and the estimation of these transformation parameters is embedded in the mixture density HMM training. This leads to a ML formulation for jointly estimating three sets of parameters, namely, the affine transformations, the mixture density means and their covariances. Following the EM principle, optimal values can be iteratively searched for by updating each parameter set while holding the two others fixed, as proposed in [5]. We now describe the complete SAT system of equations that has been considered in this study.

First, the SI means are adapted to a particular speaker r using a set of affine transformations according to :

$$\mu_k^{r \leftarrow si} = \mathbf{A}_{r,g} * \mu_k^{si} + \mathbf{b}_{r,g} \tag{1}$$

where g is the regression class of density k and $r \leftarrow si$ indicates adaptation from a speaker-independent estimation. When using the Viterbi criterion and a globally pooled covariance in the acoustic modeling, the MLLR estimation simplifies to a least mean square (LMS) approach [1] and the affine transformations can be computed by

$$[\mathbf{b}, \mathbf{A}]\,(r,g) = (\sum_{t \to g} \mathbf{o}_t^r \;\; \tilde{\mu}_{k_t}^T) * (\sum_{t \to g} \tilde{\mu}_{k_t} \; \tilde{\mu}_{k_t}^T)^{-1} \tag{2}$$

The sum $\sum_{t \to g}$ is taken over the observations $\mathbf{o}_t^r$ coming from speaker r that are assigned by Viterbi alignment to a density k_t belonging to regression class g. The vector $\tilde{\mu}_{k_t}$ stands for the augmented SI mean vector $[1, (\mu_{k_t}^{si})^T]^T$. This LMS version of MLLR has been successfully evaluated in [7] for a variety of speaker adaptation scenarios.

Second, the functional SAT model of speaker normalization can be expressed for a given density k using (1) as

$$p(\mathbf{o}_t^{si}|k) \sim \mathcal{N}(\mu_k^{si}, \Sigma_k) \implies p(\mathbf{o}_t^r|k) \sim \mathcal{N}(\mu_k^{r \leftarrow si}, \Sigma_k) \tag{3}$$

and leads to the SAT re-estimation of the means [5]:

$$\hat{\mu}_k^{si} = [\sum_{r=1}^{R} c_k^r \; \mathbf{A}_{r,g}^T \Sigma_k^{-1} \mathbf{A}_{r,g}]^{-1} * \sum_{r=1}^{R} c_k^r \; \mathbf{A}_{r,g}^T \Sigma_k^{-1} (\hat{\mu}_k^r - \mathbf{b}_{r,g}) \tag{4}$$

where c_k^r is the "count" of observations from speaker r that are assigned to density k and $\hat{\mu}_k^r$ is the speaker-dependent mean vector obtained from

$$\hat{\mu}_k^r = \frac{\sum_{t=1}^{T(r)} \gamma_k^r(t) \;\; \mathbf{o}_t^r}{c_k^r}, \quad c_k^r = \sum_{t=1}^{T(r)} \gamma_k^r(t), \tag{5}$$

$\gamma_k^r(t)$ being the posterior probability of density k for speaker r at time t. For Viterbi training, it takes the values 0 or 1.

Third, the covariance matrices are re-estimated using

$$\hat{\Sigma}_k = [\sum_{r,t}^{R,T(r)} \gamma_k^r(t) \; (\mathbf{o}_t^r - \hat{\mu}_k^{r \leftarrow si}) (\mathbf{o}_t^r - \hat{\mu}_k^{r \leftarrow si})^T] \; / \; \sum_{r=1}^{R} c_k^r \tag{6}$$

A simple interpretation of (4) can be gained by observing how the normal probability distribution is modified when the mean is subjected to an affine transformation as in (3). After straightforward algebraic manipulations we get:

$$P(\mathbf{o}_t^r|\mu_k^{r \leftarrow si}, \Sigma_k^{-1}) \sim P(\mathbf{A}_{r,g}^{-1}(\mathbf{o}_t^r - \mathbf{b}_{r,g})|\mu_k^{si}, \mathbf{A}_{r,g}^T \Sigma_k^{-1} \mathbf{A}_{r,g}), \tag{7}$$

the proportionality factor being the determinant $|\mathbf{A}_{r,g}|$. Defining the potential matrices (i.e. inverse covariances)

$$\mathbf{S}_k(r,g) \stackrel{\text{def}}{=} \mathbf{A}_{r,g}^T \Sigma_k^{-1} \mathbf{A}_{r,g}, \tag{8}$$

equation (4) can now be directly re-written as

$$\hat{\mu}_k^{si} = [\sum_{r=1}^{R} c_k^r \; \mathbf{S}_k(r,g)]^{-1} * \sum_{r=1}^{R} c_k^r \; \mathbf{S}_k(r,g) * \mathbf{A}_{r,g}^{-1} (\hat{\mu}_k^r - \mathbf{b}_{r,g}) \tag{9}$$

Hence, the SAT re-estimation of the means appears as a weighted average of the *inverse* affine transformation applied to the SD means $\hat{\mu}_k^r$, the weights being the potential matrices $\mathbf{S}_k(r,g)$ of the corresponding distributions when "moved back" to the SI acoustic space.

This suggests an (obvious) approximation: if the dependence of the potential matrices $\mathbf{S}_k(r,g)$ on the speaker r is neglected, then the matrix weights cancel out and this provides a simplified SAT mean re-estimation formula :

$$\hat{\mu}_k^{si} = \sum_{r=1}^{R} c_k^r \; \mathbf{A}_{r,g}^{-1} \; (\hat{\mu}_k^r - \mathbf{b}_{r,g}) \; / \; \sum_{r=1}^{R} c_k^r \tag{10}$$

This alternative named the "inverse transform SAT" has been investigated in [6] and also applied in [11]. Note that (10) requires the direct inversion of $\mathbf{A}_{r,g}$ which might be ill-conditioned whereas in (4) it is a weighted sum of symmetric matrices from all speakers that has to be inverted.

3. COMPUTATIONAL REQUIREMENTS AND PRACTICAL IMPLEMENTATION

In this section, we assume that the training utterances have been partitioned according to speaker and that each speaker has uttered at least a few minutes of speech to allow a robust estimation of MLLR transformations.

Beside the organization of training data, SAT introduces two main additional requirements with respect to standard HMM training. First, the MLLR algorithm has to be integrated implying some (manageable) overhead in storage and calculations, essentially matrix products and inversions. Second, the (exact) mean re-estimation leads to a considerable increase of both memory and CPU needs due to the presence of *matrix* weights in the "denominator" of (4) precluding from a straightforward implementation. Indeed, the storage volume required for accumulating these matrix counts over the speakers is Kd^2 (4 bytes) elements and since K (the total number of mixture densities) is typically in the range $2 * 10^4 \rightarrow 2 * 10^5$, this leads to a memory space of $0.1 \rightarrow 1.0$ GigaByte, for $d = 35$. As this appears rather prohibitive, another implementation consists in storing, after each speaker has been processed, the SD counts c_k^r in addition to the MLLR matrices $\mathbf{A}_{r,g}$. The denominator is "synthesized" afterwards for each density k by summing up over the speakers according to

$$\mathbf{D}_k \stackrel{\text{def}}{=} \sum_{r=1}^{R} c_k^r \; \mathbf{A}_{r,g}^T \Sigma_k^{-1} \mathbf{A}_{r,g} \tag{11}$$

However, this introduces a (strong) dependence on the number of speakers R due to the storage of the SD counts requesting $2RK$ bytes [1]. Storing the MLLR transforms requires $4RGd(d+1)$ bytes which can be more easily accommodated for, even for a large number of speakers (for 1000 speakers and 20 regression classes per speaker this amounts to about 100 MegaBytes). Nevertheless, as long as there are not more than a few hundred training speakers, this SAT implementation can be made relatively fast and flexible on a common workstation.

Concerning the re-estimation of the variances using (6),

[1] Note that these counts being sparse, they can be strongly compressed using a list organization

a whole SAT cycle can be easily performed in a single-pass if the previous updates of the SI means are combined with the latest MLLR updates to get the speaker-adapted means $\hat{\mu}_k^{r \leftarrow si}$. This keeps the overhead at a minimum without seriously slowing down the convergence.

Here follow the main steps of our one-pass SAT algorithm. Seed values for the mixture parameters come from standard SI estimations and initial MLLR transforms are computed with (2) during a preliminary pass through all training speakers' data. This provides $(\hat{\mu}_k^{si})^0$ and $\mathbf{A}_{r,g}^0, \mathbf{b}_{r,g}^0$. Each SAT iteration (indexed by i) proceeds in two parts:

- For each training speaker $r = 1, ..., R$,
  1. Transform SI means $(\hat{\mu}_k^{si})^{i-1}$ with $\mathbf{A}_{r,g}^{i-1}, \mathbf{b}_{r,g}^{i-1}$
  2. Perform Viterbi alignment using $\mu_k^{r \leftarrow si}$ to get the SD counts $(c_k^r)^i$ and SD mean vectors $\hat{\mu}_k^r$
  3. Compute new transformations $\mathbf{A}_{r,g}^i, \mathbf{b}_{r,g}^i$
  4. Accumulate the numerators of (4) and (6) for each observed density k (in 1-D arrays)
  5. Store the SD counts $(c_k^r)^i$ with $\mathbf{A}_{r,g}^i$ and $\mathbf{b}_{r,g}^i$
- For each mixture density $k = 1, ..., K$,
  1. Compute "denominator" $\mathbf{D}_k$ (11) and invert
  2. Introduce in (4) to get updated means $(\hat{\mu}_k^{si})^i$
  3. Normalize variances and mixt. weights as usual

When appling the "inverse transform SAT" (10), the matrices $\mathbf{A}_{r,g}^i$ are directly inverted in step 3 (possibly with some smoothing [6]) and the numerator of (10) is accumulated accordingly. There is no need for storing SD counts as the denominator now consists of standard SI counts.

4. DEVELOPMENT RESULTS FOR SUPERVISED BATCH ADAPTATION

First, the influence of several factors acting on the SAT estimation has been studied using WSJ0 training data and the ARPA Spoke 0 test-set of Nov'94 evaluation (20 speakers, 425 sentences, 7135 words). The base system has 3,326 tied states each with 32 mixture components. Supervised adaptation has been performed on 40 enrollment sentences per speaker using 43 phonetically derived regression classes [7]. Table 1 summarizes the results obtained with 84 training speakers, each with $\approx$10' speech.

Table 1: Development for Supervised Batch Adaptation, Training on WSJ0, Test on Spoke 0, 5K Bigram

Estimation	TRN Statist.		No Ada	Sup. Batch Adapt.		
of mixtures	Log-Lik.	Var.	WER	WER	#Hyp.	Gain
STD SI-84	29,728	.474	10.3%	8.8%	9.1k	-14.7%
SAT Uni	28,988	.448	10.8%	8.1%	6.9k	-25.4%
SAT Mul	28,225	.422	10.9%	8.0%	5.9k	-26.1%
SAT Mul-l	27,958	.414	11.7%	7.9%	5.4k	-32.5%
SAT Mul-v	28,324	.474	11.0%	8.1%	8.0k	-26.3%

Five estimation strategies have been considered, the first being standard SI training. SAT has been applied either with one single transform per speaker ('Uni') or with dynamically defined multiple regressions ('Mul'). Three SAT cycles have been performed except for the third case ('Mul-l') where seven iterations were done. The second column gives the (scaled) log-likelihood value achieved on the training data. The third column is the variance value averaged over all densities and dimensions. In the last case ('Mul-v'), the initial variances have been kept fixed, only the means being re-estimated. The word error rates (WER) are given without and after (supervised) adaptation. A measure of the decoding search cost is given by the average number of active hypotheses ('#Hyp.').

The following observations can be made :

- When more transformations as well as more SAT cycles are considered, the models show a better fit on the training data and reduced variances.
- Using SAT models "as such" without adaptation provides (increasingly) degraded results.
- The relative gains of supervised adaptation are about twice larger for the SAT models, thus entirely compensating for the worse "starting point".
- Combined with adaptation, SAT models achieve a relative improvement of 8% to 10% versus the speaker-adapted standard models.
- The search cost is reduced by as much as 40% using SAT, except for the last case with "fixed" variances.

The impact of the "inverse transform" approximation (10) is shown in table 2 hereunder, as opposed to the "exact" SAT formulation (4) for the mean re-estimation.

Table 2: Comparison of exact and inverse-transform SAT on S0 with Supervised Batch Adaptation

Estimation of SAT Means	Log-Lik. on TRN	WER No Ada	WER Sup. Ada
Uni Exact	28,988	10.82%	8.07%
Uni Approx	29,020	10.71%	8.13%
Mul Exact	28,225	10.87%	8.03%
Mul Approx	28,358	11.31%	8.37%

When using a single transformation ('Uni'), there are no significant differences. However, for multiple MLLRs there is a larger degradation mainly due, presumably, to the inaccurate inversions of some (non-smoothed) matrices.
Next, gender-dependent SAT models have been trained on WSJ0+1, using resp. 142 females and 142 males. Results on S0 are presented for supervised batch adaptation again, using a bigram or a trigram language model (LM).

Table 3: Word Error Rate (%) on Spoke 0 for Bigram WSJ0+1 Training with Supervised Adaptation

142F/142M TRN	No Ada	SUP. Ada	Rel. Gain
STD Models	7.71%	6.69%	-13.3%
SAT Models	7.93%	6.15%	-22.4%
Rel. Change	+2.9%	-8.1%	-

Table 4: Word Error Rate (%) on Spoke 0 for Trigram WSJ0+1 Training with Supervised Adaptation

142F/142M TRN	No Ada	SUP. Ada	Rel. Gain
STD Models	6.14%	5.16%	-16.0%
SAT Models	6.22%	4.72%	-24.1%
Rel. Change	+1.4%	-8.4%	-

For both LMs, a significant improvement of about 8% can be observed in tables 3 and 4, in spite of a slightly degraded "starting-point" without adaptation.

5. EXTENSION TO UNSUPERVISED INCREMENTAL ADAPTATION

As shown in the previous section, SAT models as such do not produce better results than standard ones and might even perform worse in some cases [8], unless they are (carefully) adapted to the test speaker. This can be a problem when applying unsupervised incremental adaptation (UIA). This led us to consider the possibility of exploiting the transformations jointly estimated during training to get better initial models for a new speaker.

The following algorithm suggested by ML estimation principles, has been designed much alike the selection of frequency warpings is accomplished in VTN techniques [3]. Given some (unknown) utterances of a new speaker:

1. Decode using the non adapted SAT models
2. For each speaker r considered during training:
   - Transform the SI means with $\mathbf{A}_{r,g}, \mathbf{b}_{r,g}$
   - Compute the likelihood of the decoded words
   - Select speaker r^* achieving the best likelihood
3. Transform the SI means using $\mathbf{A}_{r^*,g}, \mathbf{b}_{r^*,g}$
4. Proceed further using the transformed SAT models

To get fast unsupervised enrollment, this algorithm has been applied to just the first utterance and to be more robust against decoding errors, a single transformation has been considered per training speaker ($G=1$). Table 5 gives the S0 results using the same models as in table 3.

Table 5: Unsupervised Incremental Adaptation on S0 WSJ0+1 Training and Bigram LM

ESTIMATION of Mixtures	WER No Ada	WER Uns. Ada	Relat. Gain
STD SI-142F/M	7.71%	6.91%	-10.4%
SAT Models alone	7.93%	6.28%	-20.7%
SAT+Glob.Transf.	7.47%	6.15%	-17.6%

This leads to the following comments :

- On this set, SAT models alone perform better by 9% when undergoing UIA after each sentence.
- Using the global transform from a training speaker that best fits with the first utterance, we get an improvement of 3.1% without further adaptation.
- Combined with UIA another (slight) gain of 2% is observed and the total gain is 11% w.r. to 6.91%.

The same experiment has been run on the development set of Nov'92 (10 speakers, 410 sentences). Table 6 gives the main figures obtained with WSJ0 models. Using training transforms reduces SAT errors without adaptation and leads to a gain of 5% with UIA, from 6.86% to 6.53%.

Table 6: Unsupervised Incremental Adaptation on Dev WSJ0 Training 'SAT Mul-l' and Bigram LM

ESTIMATION of Mixtures	WER No Ada	WER Uns. Ada	Relat. Gain
STD SI-84	8.08%	6.86%	-15.1%
SAT Models alone	8.66%	6.62%	-23.5%
SAT+Glob.Transf.	8.14%	6.53%	-19.7%

Another question of interest concerns the improvements that can be achieved "asymptotically" with SAT models versus standard ones, after a large amount of speech has been processed. Therefore, we did the same experiments on the Spoke 4 test-set. This includes four US native speakers, each one having uttered 100 sentences (about 12' speech) instead of only 20 for Spoke 0. Word error rates have been computed separately for the 50 first and last sentences. Table 7 summarizes the results.

Table 7: Unsupervised Incremental Adaptation on S4 Word Error Rate for 50 first & last Sentences

ESTIMATION of Mixtures	Uns. Ada Snt 1→100	Uns. Ada Snt 1→50	Uns. Ada Snt 51→100
STD SI-142F/M	9.49%	10.37%	8.63%
SAT+Glob.Transf.	9.00%	9.54%	8.45%
Relative Change	-5.2%	-8.0%	-2.0%

It appears that the overall gain of 5.2% observed after 100 sentences is largely obtained on the first 50 sentences (8% improvement) while the last 50 contribute only marginally for 2%. This seems to indicate that SAT is particularly well suited for getting faster enrollment of new speakers.

6. CONCLUSION

Generally speaking, for supervised adaptation our results underline those reported in [5] and [8]. For unsupervised adaptation, we have shown that the enrollment of a new speaker can be sped up by selecting among the transformations estimated from the training speakers, the one that best fits with the first test utterance. More extensive tests are needed however, especially to study the influence of a mismatch between training and testing conditions, for example when a different acoustic channel is used or when the test speakers are non-native. On the other hand, MAP adaptation has been deliberately disabled in this study to be consistent with SAT estimation, although it has proven to combine very well with MLLR adaptation for standard SI models [7].

7. REFERENCES

[1] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, Vol. 9 (2), pp. 171-185, 1995.

[2] M. Padmanabhan, L.R. Bahl, D. Nahamoo, M.A. Picheny, "Speaker Clustering and Transformation for Speaker Adaptation in Large-Vocabulary Speech Recognition Systems", Proc. ICASSP, pp. 701-704, Atlanta, Georgia, USA, 1996.

[3] Li Lee and R. C. Rose, "Speaker Normalization using Efficient Frequency Warping Procedures", Proc. ICASSP, pp. 353-356, Atlanta, Georgia, USA, 1996.

[4] Alejandro Acero and Xuedong Huang, "Speaker and Gender Normalization for Continuous-Density Hidden Markov Models", Proc. ICASSP, pp. 342-345, Atlanta, Georgia, USA, 1996.

[5] Tasos Anastasakos, John McDonough, Richard Schwartz, John Makhoul, "A Compact Model for Speaker-Adaptive Training", Proc. ICSLP, pp. 1137-1140, Philadelphia, USA, October 1996.

[6] Spyros Matsoukas, Rich Schwartz, Hubert Jin, Long Nguyen, "Practical Implementations of Speaker-Adaptive Training", DARPA Speech Recognition Workshop, Chantilly, Virginia, USA, February 1997.

[7] E. Thelen, X. Aubert, P. Beyerlein, "Speaker Adaptation in the Philips System for Large Vocabulary Continuous Speech Recognition", Proc. ICASSP, pp. 1035-1038, Munich, Germany, April 1997.

[8] Tasos Anastasakos, John McDonough, John Makhoul, "Speaker Adaptive Training: A Maximum Likelihood Approach to Speaker Normalization", Proc. ICASSP, pp.1043-1046, Munich, Germany, April 97.

[9] John McDonough, Tasos Anastasakos, George Zavaliagkos, Herbert Gish, "Speaker-Adapted Training on the Switchboard Corpus", Proc. ICASSP, pp. 1059-1062, Munich, Germany, April 1997

[10] D. Pye & P. C. Woodland, "Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition", Proc. ICASSP, pp. 1047-1050, Munich, Germany, April 1997.

[11] Venkatesh Nagesha, Larry Gillick, "Studies in Transformation-Based Adaptation", Proc. ICASSP, pp. 1031-1034, Munich, Germany, April 1997.