EXPERIMENTS IN SPEAKER ADAPTATION FOR FACTOR ANALYSIS BASED SPEAKER VERIFICATION

Shou-Chun Yin[1,2], Patrick Kenny[1], Richard Rose[2]

[1] Centre de recherche informatique de Montréal (CRIM)
(shouchun.yin, pkenny)@crim.ca

[2] Department of Electrical and Computer Engineering McGill University, Montreal, Canada
(syin2,rose)@ece.mcgill.ca

ABSTRACT

This paper presents methods for supervised and unsupervised speaker adaptation of Gaussian mixture speaker models in text-independent speaker verification. The methods are based on an approach which is able to decompose speaker and channel variability so that progressive updating of speaker models can be performed while minimizing the influence of the channel variability associated with the adaptation utterances. This approach relies on a joint factor analysis model of intrinsic speaker variability and session variability where inter-session variation is assumed to result primarily from the effects of the channel [1]. These adaptation methods have been evaluated under the adaptation paradigm defined under the NIST 2005 speaker recognition evaluation plan which is based on conversational telephone speech [2]. It was found that when both target speaker model training and speaker verification trials were performed using a five minute excerpt from a single conversation, an equal error rate (EER) of 4.5% and minimum detection cost function (DCF) of 0.013 were obtained when performing unsupervised speaker adaptation during evaluation. It will be shown that this performance is comparable to that obtained by state of the art speaker verification systems that rely on a larger set of features and are trained from as many as eight conversations from the target speaker.

1. INTRODUCTION

Many of the recent advances in Gaussian mixture model (GMM) based speaker verification have come from the development of techniques for reducing the impact of sources of inter-session variability on speaker models. These include feature warping techniques designed to produce input feature vectors that have a better fit to a cummulative distribution function [3]. They also include score normalization techniques like the z-norm and t-norm which reduce the variability of the likelihood ratio scores that are used in the speaker verification decision criterion [4].

Inter-session variability is known to be particularly problematic when unsupervised speaker adaptation is used to progressively update target speaker models for each potential trial utterance during verification. Care must be taken to prevent the adaptation update for a trial utterance from simply providing a better representation of the channel characteristics associated with that particular utterance rather than improving the model representation of the speaker. This is an important and interesting problem in speaker verification because it is well known that major performance improvements can be obtained by progressive speaker adaptation [5, 6] and there is no obvious alternative way of dealing with the notorious ageing problem whereby the performance of speaker models degrades over time. Although NIST has encouraged participants in the recent annual speaker evaluation campaigns to experiment with speaker adaptation, progress has been slow and few researchers have taken up the challenge [7, 8].

The adaptation procedure presented in this paper is based on a factor analysis approach to GMM based speaker verification [1]. The important aspect of this approach from the standpoint of model adaptation is that it is based on a model that accounts for speaker and session or channel variability by two sets of latent variables called speaker factors and channel factors as summarized in Section 2. As a result of this decomposition, speaker adaptation can be performed by updating a set of speaker dependent hyperparameters as each adaptation utterance is collected and ignoring the channel effects in the various recordings. Section 3 describes how we implement this strategy. Finally, a description of the experimental study and a summary of the results for both scenarios is provided in Section 4.

Both supervised and unsupervised speaker adaptation scenarios for text-independent speaker verification are investigated in this paper. The verification tasks are based on the NIST 2005 speaker recognition evaluation plan which uses the conversational telephone speech data collected for the Mixer Corpus by the Linguistic Data Consortium [2]. The unsupervised adaptation experiments were performed on the core test data according to the specifications given in the NIST 2005 plan for this scenario [2]. This involved using a single five minute conversation to train target speaker models. Speaker verification and progressive speaker model adaptation was then performed on conversation length trial utterances from the target speaker that were randomly interspersed with imposter speaker utterances.

A supervised adaptation scenario was also implemented to provide an indication of the best performance that could be achieved under the NIST unsupervised adaptation scenario. We used the 8 conversation side test set for this experiment. We trained an initial model for each target speaker using the first enrollment utterance and updated it successively with the remaining 7 conversation sides using our progressive adaptation algorithm. We then evaluated the target speaker models obtained in this way using the test utterances provided in the 8 conversation side test set.

This work was supported by the Centre de recherche informatique de Montréal (CRIM) and by the Department of Electrical and Computer Engineering, McGill University.

2. FACTOR ANALYSIS

This section provides a brief introduction to the joint factor analysis model for GMM based speaker verification [9, 10, 11, 12]. First, the joint factor analysis model is summarized in Section 2.1 as a means for describing speaker and channel dependent GMMs using hidden variables known as speaker and channel factors. Then, Section 2.2 discusses the estimation of both the speaker-independent and speaker-dependent hyperparameter sets that form the factor analysis model.

2.1. Speaker factors and channel factors

Gaussian mixture models have become the most commonly used representation for text-independent speaker recognition and speaker verification. The work described in this paper relies on a GMM based speaker verification system where speakers are represented by the means, covariance matrices, and weights of a mixture of C multivariate diagonal-covariance Gaussian densities defined over an F dimensional feature space. The GMM parameters for a particular target speaker are estimated by adapting the parameters of a universal background model (UBM), which is a C component GMM trained from a large speaker population, using utterances from the target speaker.

Assuming that a C component GMM in an F dimensional feature space is used to characterize a speaker s, it is convenient to describe the speaker by concatenating the GMM mean vectors into a CF dimensional super vector which we denote by $\boldsymbol{s}$. In order to incorporate channel effects into the model, allowing for the fact that there will be many utterances from speaker s taken from many different channels, we will use the notation $\boldsymbol{M}$ to refer a speaker and channel dependent super vector. We assume that $\boldsymbol{M}$ can be decomposed into speaker-dependent and channel-dependent super vectors,

$$\boldsymbol{M} = \boldsymbol{s} + \boldsymbol{c}. \tag{1}$$

In Equation 1, $\boldsymbol{s}$ is the speaker-dependent super vector which is independent of session variations and $\boldsymbol{c}$ is a channel-dependent super vector. Both $\boldsymbol{s}$ and $\boldsymbol{c}$ are assumed to be normally distributed.

Kenny et al. have described a factor analysis model for speaker verification where both the speaker and channel and super vectors can be represented in separate low dimensional subspaces [11]. A simplified version of this model is used here to facilitate speaker adaptation within the factor analysis framework for speaker verification. The simplified factor analysis model assumes that only the channel-dependent super vector is represented in a low dimensional channel space. The speaker-dependent super vector is represented as

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{d}\boldsymbol{z}. \tag{2}$$

where the hyperparameters $\boldsymbol{m}$ and $\boldsymbol{d}$ are estimated from a large ancillary training set and and $\boldsymbol{z}$ is speaker-dependent random vector assumed to have a standard normal distribution. In Equation 2, $\boldsymbol{m}$ is the speaker-independent super vector. In our case, the ancillary data corresponds to the concatenated mean vectors of the universal background model (UBM) that is trained from a large population of "background speakers." If enrollment data for a speaker is given and channel effects are ignored, then a point estimate of $\boldsymbol{z}$ and hence of $\boldsymbol{s}$ can be obtained from the enrollment data by classical MAP estimation.

The super vector $\boldsymbol{c}$ which represents the channel effects in an utterance, is assumed to be distributed according to

$$\boldsymbol{c} = \boldsymbol{u}\boldsymbol{x} \tag{3}$$

where $\boldsymbol{u}$ is a rectangular matrix of low rank and $\boldsymbol{x}$ has a standard normal distribution. The entries of $\boldsymbol{x}$ are known as channel factors. This is equivalent to saying that $\boldsymbol{c}$ is normally distributed with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{u}\boldsymbol{u}^*$. Given an utterance by a speaker whose super vector $\boldsymbol{s}$ is known, a point estimate of $\boldsymbol{x}$ and hence of $\boldsymbol{c}$ can be obtained by eigenchannel MAP estimation.

In practice, channel effects cannot be ignored in estimating $\boldsymbol{s}$ from Equation 1 and $\boldsymbol{s}$ is not known in estimating $\boldsymbol{c}$. To get around this difficulty, a Gauss-Seidel type iterative procedure is proposed in [13]. The solution used here involves calculating the joint posterior distribution of the hidden variables $\boldsymbol{z}$ and $\boldsymbol{x}$ [10], and is briefly described in the appendix. In order to fully specify the model, we also need a residual diagonal covariance matrix $\boldsymbol{\Sigma}$ (of dimension $CF \times CF$) whose role is to model the variability which is not not captured by $\boldsymbol{s}$ and $\boldsymbol{c}$.

2.2. Speaker-independent and Speaker-dependent hyperparameter estimation

The hyperparameters $\boldsymbol{m}, \boldsymbol{u}$ and $\boldsymbol{d}$ model the prior distribution of a GMM super vector $\boldsymbol{M}$: $\boldsymbol{s}$ is normally distributed with expectation $\boldsymbol{m}$ and covariance matrix $\boldsymbol{d}^2$, and $\boldsymbol{c}$ is normally distributed with expectation zero and covariance matrix $\boldsymbol{u}\boldsymbol{u}^*$. We will refer to them as *speaker-independent* hyperparameters.

In order to describe our progressive adaptation algorithm we have to introduce speaker-dependent hyperparameters $\boldsymbol{m}(s)$ and $\boldsymbol{d}(s)$ which are estimated using some enrollment data for a speaker s and used to model the posterior distribution of a speaker-specific super vector $\boldsymbol{s}$. The assumption is that

$$\boldsymbol{s} = \boldsymbol{m}(s) + \boldsymbol{d}(s)\boldsymbol{z}. \tag{4}$$

Thus in the posterior distribution, we assume that $\boldsymbol{s}$ is assumed normally distributed with expectation $\boldsymbol{m}(s)$ and covariance matrix $\boldsymbol{d}^2(s)$. Whereas in Equation 2, $\boldsymbol{d}$ models the variability of the speaker population as a whole, $\boldsymbol{d}(s)$ models the residual uncertainty in the point estimate of $\boldsymbol{s}$ that arises from the fact that the enrollment data is of limited duration.

In order to estimate the speaker-independent hyperparameter set ($\boldsymbol{m}, \boldsymbol{u}, \boldsymbol{d}$ and $\boldsymbol{\Sigma}$) or $\boldsymbol{\Lambda}$ for short, we use a large ancillary training set and the EM algorithms described in [11, 14]. The training data set is based on: the LDC releases of Switchboard II, Phases 1, 2 and 3; Switchboard Cellular, Parts 1 and 2; the Fisher English Corpus, Part 1 and the NIST 2004 evaluation data. We chose only those speakers for whom multiple recordings (6 or more) were available in order to model channel variability properly. The female training data contains 612 speakers and 6764 conversation sides, and the male training data contains 463 speakers and 5254 conversation sides.

In estimating the speaker-independent hyperparameters, we skipped the 'adaptation to the target speaker population' step in the Section 3 of [11], in order to follow the NIST evaluation protocol. However, this step *applied in the case of a single speaker* is the fundamental idea used to estimate the speaker-dependent hyperparameters $\boldsymbol{\Lambda}(s)$ for each target speaker s. Simply stated, the initial hyperparameters come from $\boldsymbol{\Lambda}$. We fix the speaker-independent $\boldsymbol{u}$ and $\boldsymbol{\Sigma}$, and re-estimate the speaker-dependent $\boldsymbol{m}$ and $\boldsymbol{d}$ using the incoming utterance of the target speaker s.

The likelihood function that we use to make verification decisions and the enrollment procedure which we use to estimate the posterior distribution of a target speaker's super vector $\boldsymbol{s}$ are the same as in [12]. The posterior calculation needed for enrollment is summarized in the appendix. The only difference is that in progressive speaker adaptation, we use the speaker-dependent hyperparameters as the starting point for enrolling a target speaker rather than the speaker-independent hyperparameters (as in the non-adaptive case [12].

3. PROGRESSIVE SPEAKER ADAPTATION

The algorithm used for progressive speaker adaptation is the speaker dependent hyperparameter estimation algorithm from [10]. A speaker dependent hyperparameter set $\boldsymbol{\Lambda}(s)$ is updated for the factor analysis model, whenever a new utterance by the speaker becomes available. The algorithm is summarized in the following theorem which is a special case of Theorem 10 in [10]. The likelihood function $P_{\boldsymbol{\Lambda}}$ in the statement of this theorem is the factor analysis likelihood function defined in [10].

Theorem: Suppose we are given a speaker s, a hyperparameter set $\boldsymbol{\Lambda}_0$ *where* $\boldsymbol{\Lambda}_0 = (\boldsymbol{m}_0, \boldsymbol{u}_0, \boldsymbol{d}_0, \boldsymbol{\Sigma}_0)$ *and a recording* $\mathcal{X}$*. Let* $\boldsymbol{\Lambda}(s)$ *be the hyperparameter set* $(\boldsymbol{m}(s), \boldsymbol{u}_0, \boldsymbol{d}(s), \boldsymbol{\Sigma}_0)$ *where*

$$\begin{aligned} \boldsymbol{m}(s) &= \boldsymbol{m}_0 + \boldsymbol{d}_0 \boldsymbol{\mu}_{\boldsymbol{z}}(s) \\ \boldsymbol{d}(s) &= \boldsymbol{d}_0 \boldsymbol{K}_{\boldsymbol{zz}}^{1/2}(s) \end{aligned} \tag{5}$$

and $\boldsymbol{\mu}_{\boldsymbol{z}}(s)$ *and* $\boldsymbol{K}_{\boldsymbol{zz}}(s)$ *are the posterior expectation and covariance of* $\boldsymbol{z}$ *calculated using* $\boldsymbol{\Lambda}_0$ *(as explained in the appendix). Then* $P_{\boldsymbol{\Lambda}(s)}(\mathcal{X}) \geq P_{\boldsymbol{\Lambda}_0}(\mathcal{X})$.

Taking $\boldsymbol{\Lambda}_0(s)$ to the speaker-independent hyperparameter set and applying this recursively we obtain a sequence of speaker dependent hyperparameter sets $\boldsymbol{\Lambda}_1(s), \boldsymbol{\Lambda}_2(s), \ldots$ as follows:

$$\begin{aligned} \boldsymbol{m}_i(s) &= \boldsymbol{m}_{i-1}(s) + \boldsymbol{d}_{i-1}(s) \boldsymbol{\mu}_{\boldsymbol{z}} \\ \boldsymbol{d}_i(s) &= \boldsymbol{d}_{i-1}(s) \boldsymbol{K}_{\boldsymbol{zz}}^{1/2} \end{aligned} \tag{6}$$

where the posterior expectations and covariances are calculated using $\boldsymbol{\Lambda}_{i-1}(s)$, for $i = 1, 2, \ldots$. For each iteration i, $\boldsymbol{m}_i(s)$ represents the estimate of the speaker's super vector after the ith recording of the speaker has been collected and $\boldsymbol{d}_i^2(s)$ represents the uncertainty in this estimate. It can easily be shown that $\boldsymbol{d}_i^2(s) \rightarrow \boldsymbol{0}$ as the amount of adaptation tends to infinity, as one would expect.

The key practical issue in using this algorithm in unsupervised speaker adaptation is to decide when to update a speaker model using a given utterance. For a given task, the optimal threshold for this may not be the same as the threshold which optimizes the NIST detection cost function for the task as a whole!

Another problem that arises is in how to apply t-norm in situations where the models for imposter speakers have been estimated using varying numbers of recordings (rather than with a fixed number of recordings such as 1 or 8 as in traditional NIST evaluations). Our experience has been that a straightforward implementation of t-norm does not give good results in this situation when it is used in conjunction with the likelihood function that we use for making verification decisions. The problem might come from a side-effect of progressive adaptation which has previously been reported in text-dependent speaker verification [5, 6], namely that verification scores tend to drift as the amount of adaptation data increases which makes it difficult to normalize them. Our principal motivation for experimenting with the supervised adaptation scenario was to avoid having to confront this problem at the outset.

4. EXPERIMENTS

The unsupervised adaptation results that we report were obtained on the core condition of the NIST 2005 evaluation using all of the trials in this condition rather than the 'common' subset [2] and the supervised adaptation results were obtained on 8 conversation side condition. For the core condition we carried out in the trials by following the training and test utterance designations given by NIST.

In all experiments, gender-dependent UBM's were used with 2048 Gaussians and 26 dimension acoustic feature vectors consisted of 13 Gaussianized cepstral features and their first derivatives. Both equal error rates (EER) and the minimum values of the NIST detection cost function (DCF) are reported where both were obtained in both supervised and unsupervised speaker adaptation scenarios.

4.1. Supervised speaker adaptation

In our supervised speaker adaptation experiments we used the data from the '8 conversation 2-channel' condition of the NIST 2005 evaluation, with a total of 2230 target trials and 21,216 non-target trials. The first conversation side for each speaker s is used to obtain an initial estimate of speaker-dependent hyperparameter set $\boldsymbol{\Lambda}(s)$ as described in Section 3, then the remaining conversation sides are used to update $\boldsymbol{\Lambda}(s)$ for seven iterations before performing the prescribed verification tests. A combination of the z-norm and t-norm, called zt-norm, was used to normalize the likelihood ratio scores [13].

Channel Factors	Adaptation	EER	DCF
25	supervised	3.39%	0.0113
25	non-adaptive	6.9%	0.0221
100	supervised	3.12%	0.0099
100	non-adaptive	6.44%	0.0193

Table 1. *Results on the '8 conversation 2 channel' test set obtained without adaptation (1 enrollment utterance) and with supervised speaker adaptation (8 enrollment utterances).* zt*-norm score normalization. All trials, male and female.*

Table 1 shows the results for supervised speaker adaptation in terms of both EER and DCF. Both systems shown in the table use zt score normalization. Using 100 channel factors gave an EER of 3.12% and a minimum value of DCF of 0.0099. Considering that we use only short term acoustic features in our system, these are very interesting results. For comparison, the best results obtained in the 2005 NIST Speaker Recognition Evaluation using 8 conversation sides, namely an EER of 3.02% and 'hard decision' DCF of 0.0097 (all trials, male and female), was obtained by the SRI-2 system.

The main point of Table 1 is to compare the results obtained with and without using speaker adaptation. The 'non-adaptive' results in the table refer to the case where a single utterance is used to enroll each target speaker. For the same numbers of channel factors using the same zt score normalization, both EER and DCF drop by half as a result of supervised adaptation. This shows that the hyperparameter update algorithm described in Section 3 has

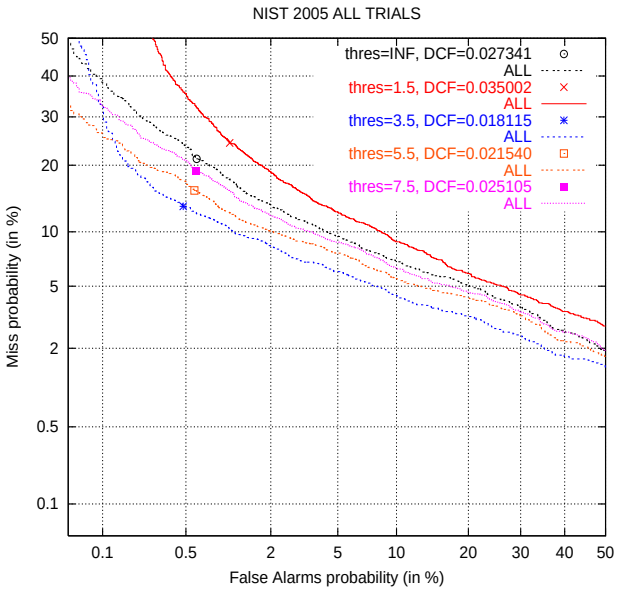

Fig. 1. *DET curves for various adaptation thresholds used in the unsupervised speaker adaptation. Joint factor analysis with 25 channel factors. z-norm score normalization. All trials, male and female.*

been properly implemented and is potentially very effective for unsupervised speaker adaptation.

4.2. Unsupervised speaker adaptation

We used the core condition data ('1 conversation 2-channel') in the NIST 2005 evaluation data as the test set for our unsupervised speaker adaptation experiments. This consists of 2771 target trials (1231 for male trials and 1540 for female trials) and 28, 472 non-target trials (12, 317 for male trials and 16, 155 for female trials). The verification scores for all trials are calculated and the normalized scores are compared with a common threshold which we will refer to as the *adaptation* threshold. If the score is higher than the adaptation threshold, the corresponding trial is allowed for speaker adaptation before performing the next trial. This threshold is not the same as the threshold which is used to minimize the value of the NIST detection cost function and which we will refer to as the *decision* threshold; rather each value of the adaptation threshold results in a different DET curve as shown in Fig. 1. The trials for each target speaker are performed in a fixed order given by NIST to follow the NIST evaluation protocol. DET curves in both unsupervised adaptation mode and in non-adaptive mode (adaptation threshold = ∞) are shown and the results are summarized in Table 2.

These results were obtained with 25 channel factors and z-norm score normalization. Note that decreasing the adaptation threshold, which increases the amount of enrollment data for target speakers, initially improves the values for both EER and DCF. However, Table 2 also shows that the performance degrades once the the adaptation threshold becomes low enough so that a significant number of non-target utterances are used for adaptation. Table 3 shows the number of target and non-target utterances that are used for speaker adaptation as the adaptation threshold is varied over a range from 1.5 to ∞.

Adaptation Threshold	EER	DCF
∞	7.5%	0.027
7.5	7.2%	0.025
5.5	6.5%	0.022
3.5	5.9%	0.018
1.5	9.2%	0.035

Table 2. *Performance for various adaptation thresholds used in unsupervised speaker adaptation. Joint factor analysis with 25 channel factors. z-norm score normalization. All trials, male and female.*

Adaptation Threshold	No. targets accepted	No. non-targets accepted
∞	0	0
7.5	966	4
5.5	1764	10
3.5	2443	205
1.5	2565	4066

Table 3. *Comparison of the number of target and non-target utterances accepted for speaker adaptation using different values of adaptation threshold. All trials, male and female. Joint factor analysis with 25 channel factors. z-norm score normalization.*

The best adaptation threshold value we found used for all trials was 3.5 which provides an EER of 5.9% and a minimum DCF of 0.018. This performance may perhaps represent a slightly optimistic estimate of the performance obtained for a priori threshold settings. In any speaker verification system using progressive speaker adaptation, two thresholds have to be determined. These are the adaptation threshold which determines when an utterance will be used to update a target speaker model and the decision threshold which determines how to make a hard decision about whether or not to reject a claimant. There is no *a priori* reason to believe that these two thresholds should be the same because the decision threshold is largely determined by the parameters which specify the NIST detection cost function.

Adaptation	Normalization	EER	DCF
supervised	zt-norm	3.4%	0.011
unsupervised	z-norm	5.9%	0.018

Table 4. *Supervised speaker adaptation results using zt-norm score normalization and the best unsupervised speaker adaptation results using an adaptation threshold 3.5 and z-norm score normalization. Joint factor analysis with 25 channel factors. All trials, male and female.*

Results obtained with supervised adaptation and 25 channel factors of our supervised test set (i.e. '8 conversation 2 channel') are given in Table 4. They are better than the results obtained in our unsupervised experiment but this is only to be expected since supervised adaptation is bound to perform better than unsupervised.

This comparison also raises the question of whether t-norm or zt-norm might prove useful in speaker adaptation. If t-norm or zt-norm is applied for the results obtained using an adaptation threshold of 3.5, t-norm does not work well as discussed in Section 3; however, zt-norm still gives good results with an EER of 4.5% and a minimum DCF of 0.013. It is interesting since zt-norm combines the idea of z-norm and t-norm by implementing z-norm followed by t-norm [13], and there is no phenomenon of an obvious abnormal drift in verification scores as occurred in the case of t-norm. One possible reason might come from z-norm which compensates the side effect of progressive adaptation in the situation using zt-norm. A comparison for all the speakers in the test set that we used for our unsupervised adaptation experiments is given in Table 5. (Note that for the non-adaptive results, there is just one enrollment utterance per speaker in this case rather than eight.) In order to make a fair comparison with supervised speaker adaptation, perhaps the unsupervised adaptation results using a large number of channel factors, such as 100, should be generated.

Adaptation	Normalization	EER	DCF
unsupervised	z-norm	5.9%	0.018
unsupervised	zt-norm	4.5%	0.013
non-adaptive	z-norm	7.8%	0.027
non-adaptive	zt-norm	6.9%	0.022

Table 5. *Joint factor analysis with 25 channel factors. All trials, male and female. Non-adaptive results (obtained with a single enrollment utterance) and the best unsupervised speaker adaptation results obtained using an adaptation threshold of 3.5.*

5. DISCUSSION

The main contribution of this paper is to propose an effective way to do speaker adaptation in factor analysis based speaker verification. In our supervised adaptation experiments we showed that substantial improvements in performance could be obtained by implementing our progressive adaptation algorithm with 8 utterances per speaker when compared with using a single enrollment utterance. Both equal error rates (EER) and minimum values of the NIST detection cost function (DCF) can be reduced by half as reported in Table 1. We obtained our best results using 100 channel factors: an EER of 3.12% and a minimum detection cost of 0.0099 which compares favorably with state of the art results on the NIST extended data tasks.

In the unsupervised speaker adaptation experiments, we obtained an EER of 5.9% and a minimum DCF of 0.018 using z-norm and an EER of 4.5% and a minimum DCF of 0.013 using zt-norm, a substantially better result than we were able to achieve with non-adaptive speaker verification on the same data set.

Perhaps the most important open question concerns score normalization. Currently we just use the z and zt score normalization for unsupervised speaker adaptation and we are investigating the use of t-norm. We hope that techniques such as Frame Count Dependent Thresholding will prove to be effective [6] in t-norm, and further improve the performance using zt-norm. If the side effect of unsupervised progressive adaptation using t-norm comes from different amount of adaptation data for each speaker model, a new idea of adaptive t-norm is worth investigating which adapts the t-norm models as well as adapting the speaker models during verification.

6. REFERENCES

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," submitted to *IEEE Trans. Audio, Speech, and Language Processing*. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[2] (2005) The NIST year 2005 speaker recognition evaluation plan. [Online]. Available: http://www.itl.nist.gov/iad/894.01/tests/spk/2005

[3] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, Crete, Greece, June 2001.

[4] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–52, 2000.

[5] N. Mirghafori and L. Heck, "An adaptive speaker verification system with speaker dependent a priori decision thresholds," in *Proc. ICSLP*, Denver, Colorado, Sept. 2002.

[6] N. Mirghafori and M. Hébert, "Parameterization of the score threshold for a text-dependent adaptive speaker verification system," in *Proc. ICASSP 2004*, Montreal, Canada, May 2004.

[7] D. A. van Leeuwen, "Speaker adaptation in the NIST speaker recognition evaluation," in *Proc. Eurospeech*, Lisbon, Portugal, Sept. 2005.

[8] C. Barras, S. Meignier, and J.-L. Gauvain, "Unsupervised online adaptation for speaker verification over the telephone," in *Proc. Odyssey 2004*, Toledo, Spain, June 2004.

[9] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, May 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[10] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and alogorithms." [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[11] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. ICASSP 2005*, Philadelphia, PA, Mar. 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

[12] ——, "Improvements in factor analysis based speaker verification," in *Proc. ICASSP 2006*, Toulouse, France, May 2006.

[13] R. Vogt, B. Baker, and S. Sridharan, "Modeling session variability in text-independent speaker verification," in *Proc. Eurospeech*, Lisbon, Portugal, Sept. 2005.

[14] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," submitted to *IEEE Trans. Speech Audio Processing*. [Online]. Available: http://www.crim.ca/perso/patrick.kenny/

A. APPENDIX

In this appendix we summarize the calculations needed to evaluate the joint posterior distribution of the hidden variables $\boldsymbol{x}$ and $\boldsymbol{z}$ given a single utterance for a speaker. First some notation. For each mixture component c, let N_c be the total number of observation vectors in the utterance for the given mixture component and set

$$F_c = \sum_t X_t \tag{7}$$

where the sum extends over all observations X_t aligned with the given mixture component. Let $\boldsymbol{N}$ be the $CF \times CF$ diagonal matrix whose diagonal blocks are $N_c I$ (for $c = 1, \ldots, C$) where I is the $F \times F$ identity matrix. Let $\boldsymbol{F}$ be the $CF \times 1$ vector obtained by concatenating F_c (for $c = 1, \ldots, C$). If

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{z} \end{pmatrix}$$

then the posterior distribution of $\boldsymbol{X}$ is Gaussian of the same form as the posterior distribution described in Proposition 1 of [9]. Specifically, if $\boldsymbol{V}$ and $\boldsymbol{L}$ are the matrices defined by

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{u} & \boldsymbol{d} \end{pmatrix} \tag{8}$$

$$\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{V}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{V}. \tag{9}$$

then the posterior distribution of $\boldsymbol{X}$ has covariance matrix $\boldsymbol{L}^{-1}$ and mean $\boldsymbol{L}^{-1} \boldsymbol{V}^* \boldsymbol{\Sigma}^{-1} (\boldsymbol{F} - \boldsymbol{N}\boldsymbol{m})$. Thus calculating the posterior distribution of $\boldsymbol{X}$ is essentially a matter of inverting the matrix $\boldsymbol{L}$.

A straightforward calculation shows that $\boldsymbol{L}$ can be written as

$$\begin{pmatrix} \boldsymbol{I} + \boldsymbol{u}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{u} & \boldsymbol{u}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{d} \\ \boldsymbol{d} \boldsymbol{N} \boldsymbol{\Sigma}^{-1} \boldsymbol{u} & \boldsymbol{I} + \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{d}^2 \end{pmatrix}. \tag{10}$$

So $\boldsymbol{L}^{-1}$ can be calculated by using the identity

$$\begin{pmatrix} \boldsymbol{\alpha} & \boldsymbol{\beta} \\ \boldsymbol{\beta}^* & \boldsymbol{\gamma} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\zeta}^{-1} & -\boldsymbol{\zeta}^{-1} \boldsymbol{\beta} \boldsymbol{\gamma}^{-1} \\ -\boldsymbol{\gamma}^{-1} \boldsymbol{\beta}^* \boldsymbol{\zeta}^{-1} & \boldsymbol{\gamma}^{-1} + \boldsymbol{\gamma}^{-1} \boldsymbol{\beta}^* \boldsymbol{\zeta}^{-1} \boldsymbol{\beta} \boldsymbol{\gamma}^{-1} \end{pmatrix}$$

where

$$\boldsymbol{\zeta} = \boldsymbol{\alpha} - \boldsymbol{\beta} \boldsymbol{\gamma}^{-1} \boldsymbol{\beta}^*$$

with

$$\begin{aligned} \boldsymbol{\alpha} &= \boldsymbol{I} + \boldsymbol{u}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{u} \\ \boldsymbol{\beta} &= \boldsymbol{u}^* \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{d} \\ \text{and } \boldsymbol{\gamma} &= \boldsymbol{I} + \boldsymbol{\Sigma}^{-1} \boldsymbol{N} \boldsymbol{d}^2. \end{aligned}$$

Of course, since the dimensions of $\boldsymbol{L}$ are enormous (namely $(CF + R_C) \times (CF + R_C)$ where R_C is the rank of $\boldsymbol{u}$), care has to be taken to evaluate only those entries of $\boldsymbol{L}^{-1}$ which are actually needed.