The Role of Speaker Factors in the NIST Extended Data Task

Patrick Kenny[1], *Najim Dehak*[1,2], *Réda Dehak*[3], *Vishwa Gupta*[1] *and Pierre Dumouchel*[1,2]

[1]Centre de recherche informatique de Montréal (CRIM), Montréal, Canada
[2]École de Technologie Supérieure (ETS), Montréal, Canada
[3]Laboratoire de Recherche et de Développement de l'EPITA (LRDE), Paris, France
{patrick.kenny, najim.dehak, vishwa.gupta, pierre.dumouchel}@crim.ca
reda.dehak@lrde.epita.fr

Abstract

We tested factor analysis models having various numbers of speaker factors on the core condition and the extended data condition of the 2006 NIST speaker recognition evaluation. In order to ensure strict disjointness between training and test sets, the factor analysis models were trained without using any of the data made available for the 2005 evaluation. The factor analysis training set consisted primarily of Switchboard data and so was to some degree mismatched with the 2006 test data (drawn from the Mixer collection). Consequently, our initial results were not as good as those submitted for the 2006 evaluation. However we found that we could compensate for this by a simple modification to our score normalization strategy, namely by using 1000 z-norm utterances in zt-norm.

Our purpose in varying the number of speaker factors was to evaluate the eigenvoice MAP and classical MAP components of the inter-speaker variability model in factor analysis. We found that on the core condition (i.e. 2–3 minutes of enrollment data), only the eigenvoice MAP component plays a useful role. On the other hand, on the extended data condition (i.e. 15–20 minutes of enrollment data) both the classical MAP component and the eigenvoice component proved to be useful provided that the number of speaker factors was limited. Our best result on the extended data condition (all trials) was an equal error rate of 2.2% and a detection cost of 0.011.

1. Introduction

Classical MAP adaptation [1, 2] is by far the most popular type of speaker modeling in text independent speaker recognition but our experience has been that MAP adaptation using speaker factors [3, 4, 5] is generally more effective, at least in situations where limited amounts of enrollment data are available. The reason for this is that factor analysis is capable of explaining most inter-speaker variability using a relatively small number of hidden variables, so that only a small number of free parameters need to be estimated at enrollment time (contrary to classical MAP). However if large amounts of enrollment data are available, the assumption that speaker variability can be accounted for in such an economical way could prove to be harmful. Our principal purpose in this paper is to explore this question in the context of the extended data condition of the NIST speaker recognition evaluations (SRE's) where 15–20 minutes of enrollment data are available for each target speaker.

In the 2006 NIST SRE we reported an equal error rate (EER) of 1.7% and a detection cost function (DCF) value of 0.009 on the English language trials of the extended data condition. We suspected that these results were too good to be true considering that they were obtained with a stand-alone system using only short term acoustic features and suggested that they may be attributable in part to the fact that NIST had recycled some of the 2005 data for the 2006 evaluation and we had used the 2005 data in training our factor analysis models. (In [6] we reported how failing to keep the training and test sets disjoint could produce extremely misleading results.)

Thus in performing the experiments reported here we were careful to exclude the 2005 SRE data from our factor analysis training set. As a result, the proportion of Mixer data in our training set decreased from 50% to 20%. This introduced a mismatch between our training set and the 2006 test set which we knew from previous experience [4] was likely to adversely affect the performance of the factor analysis model. However we found that we could compensate for this in large measure by increasing the dimension of our acoustic feature vectors from 26 to 40 and by substantially increasing the number of imposters used for score normalization. (In particular, we found it helpful to use 1000 z-norm imposter utterances for each gender.) Finally, by optimizing the number of speaker factors used in the extended data condition, we found that we could obtain results almost identical to those submitted in 2006 (an EER of 1.9% and a DCF of 0.010).

2. Two ways of modeling inter-speaker variability

Joint factor analysis is a model of speaker and session variability in Gaussian mixture models (GMM's). We have described elsewhere how we estimate the hyperparameters that specify a factor analysis model and how we use it for speaker verification [4, 5]. In order to formulate precisely the problem that we address in this paper we will begin by recapitulating the basic assumptions in factor analysis.

Let C be the number of components in the universal background model and F the dimension of the acoustic feature vectors. We use the term supervector to refer to the CF dimensional vector obtained by concatenating the F dimensional mean vectors in the GMM corresponding to a given utterance.

Our assumptions are as follows. Firstly we assume that a speaker and channel-dependent supervector $\boldsymbol{M}$ can be decomposed into a sum of two supervectors, a speaker supervector $\boldsymbol{s}$ and a channel supervector $\boldsymbol{c}$:

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

where $\boldsymbol{s}$ and $\boldsymbol{c}$ are statistically independent and normally distributed.

Secondly, we assume that the distribution of $\boldsymbol{c}$ has a hidden variable description of the form

$$\boldsymbol{c} = \boldsymbol{u}\boldsymbol{x} \tag{2}$$

where $\boldsymbol{u}$ is a rectangular matrix of low rank and $\boldsymbol{x}$ is a normally distributed random vector. We refer to the components of $\boldsymbol{x}$ as channel factors. (This assumption is equivalent to saying that $\boldsymbol{c}$ is normally distributed with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{u}\boldsymbol{u}^*$.)

Thirdly, we assume that the distribution of $\boldsymbol{s}$ has a hidden variable description of the form

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y} + \boldsymbol{d}\boldsymbol{z} \tag{3}$$

where $\boldsymbol{m}$ is a $CF \times 1$ supervector; $\boldsymbol{v}$ is a rectangular matrix of low rank and $\boldsymbol{y}$ is a normally distributed random vector whose components are referred to as speaker factors; $\boldsymbol{d}$ is a $CF \times CF$ diagonal matrix and $\boldsymbol{z}$ is a normally distributed CF dimensional random vector. (This assumption is equivalent to saying that $\boldsymbol{s}$ is normally distributed with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{d}^{\boldsymbol{2}} + \boldsymbol{v}\boldsymbol{v}^*$.) Our concern in this paper is with the relative importance of the terms $\boldsymbol{v}\boldsymbol{y}$ and $\boldsymbol{d}\boldsymbol{z}$ in (3).

If $\boldsymbol{v} = \boldsymbol{0}$ and $\boldsymbol{u} = \boldsymbol{0}$ then our third assumption is the same as in classical MAP [1]; on the other hand if $\boldsymbol{d} = \boldsymbol{0}$ and $\boldsymbol{u} = \boldsymbol{0}$ the assumption is the same as in eigenvoice MAP [7]. In the latter case, the speaker supervector s is constrained to lie in the affine space $\boldsymbol{m} +$ Range $(\boldsymbol{v}\boldsymbol{v}^*)$ (we refer to this as the speaker space) but no such constraint is imposed in classical MAP.

Classical MAP adaptation can only adapt those Gaussians which are seen in the enrollment data but, if large amounts of enrollment data are available, the subspace constraint may be a hindrance in getting a good estimate of the speaker supervector $\boldsymbol{s}$.

On the other hand the subspace constraint is helpful if only small amounts of enrollment data are available, since only a small number of free parameters need to be estimated at enrollment time. The fact that the supervector covariance matrix is full rather than diagonal in this case ensures that MAP adaptation takes account of the correlations between the different Gaussians in a speaker supervector so that all of the Gaussians are updated at enrollment time even if only a small fraction of them are observed.

An extreme example of the effectiveness of speaker factors can be found in [8] which is concerned with the use of factor analysis to model syllable-level prosodic features. The number of feature vectors per conversation side is only about 400; it is unrealistic to expect classical MAP adaptation to be very effective in this situation. In working with the core condition of the 2005 evaluation we also found the most effective method of speaker modeling was to use a large number (300) of speaker factors [5].

Much of the mathematical complexity in [3] is a result of including the term $\boldsymbol{d}\boldsymbol{z}$ in (3); on the other hand essentially no simplification can be achieved by setting $\boldsymbol{v} = \boldsymbol{0}$. Thus it is very tempting to suppress the term $\boldsymbol{d}\boldsymbol{z}$ altogether but it would be premature to do so until we have determined whether there any situations in which it might prove to be effective. This is the motivation for the experiments in this paper

3. Experimental setup

In this section we describe the set up common to all of our experiments.

3.1. Enrollment and test data

Our experiments are carried out on the core condition and the 8 conversation training condition (also known as the extended data condition) of the NIST 2006 speaker recognition evaluation (SRE) [9].

For the core condition, there were 350 male and 461 female target speakers and there were and 51,448 test utterances. For the 8 conversation training condition, there were 298 male and 402 female target speakers and 32,509 test utterances.

3.2. Feature Extraction

We used 19 cepstral coefficients rather than 12 as in our previous work.

These coefficients together with a log energy feature were extracted using a 25 ms Hamming window and a 10 ms frame advance and they were subjected to feature warping [10] using a 3 s sliding window. Delta coefficients were calculated using a 5 frame window giving a total of 40 features.

3.3. Factor analysis training data

We trained 2 gender dependent universal background models (UBM's) having 1024 Gaussians and several gender dependent factor analysis models using the algorithms described in [5]. These factor analysis models differed from each other as regards the number of speaker factors but the number of channel factors was fixed at 50 in all cases.

For training UBM's we used Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; the Fisher English Corpus, Part 1; the NIST 2003 Language recognition evaluation data set; and the NIST 2004 SRE enrollment and test data. (About 200 hours of speech data for each gender.)

For training factor analysis models we used the LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; and the NIST 2004 SRE data. (For each speaker with more than 5 recordings, we used all of the recordings of the speaker for factor analysis training.)

3.4. Imposters

The verification decision scores obtained with the factor analysis models were normalized using zt-norm. We used 283 t-norm speakers in the female case and 227 in the male case. We used 1000 z-norm utterances for each gender. The imposter utterances and speakers were chosen randomly from the factor analysis training data. Our motivation for using such a large number of z-norm utterances was to test the top-norm method proposed by Zigel and Wasserblat in [11].

3.5. Reference systems

We use the results submitted at the time of the NIST 2006 SRE as a reference [12]. These results were obtained using gender dependent factor analysis models having 300 speaker factors and 75 channel factors. The UBM's had 2048 Gaussians and we used 12 cepstral coefficients rather than 19.

These UBM's and factor analysis models were trained on essentially the same data as for the experiments in the present paper with one exception: the NIST 2005 SRE enrollment and test data were included in the training set used to obtain the reference results. (Note that 2005 SRE data was not included in list given in Section 3.3).

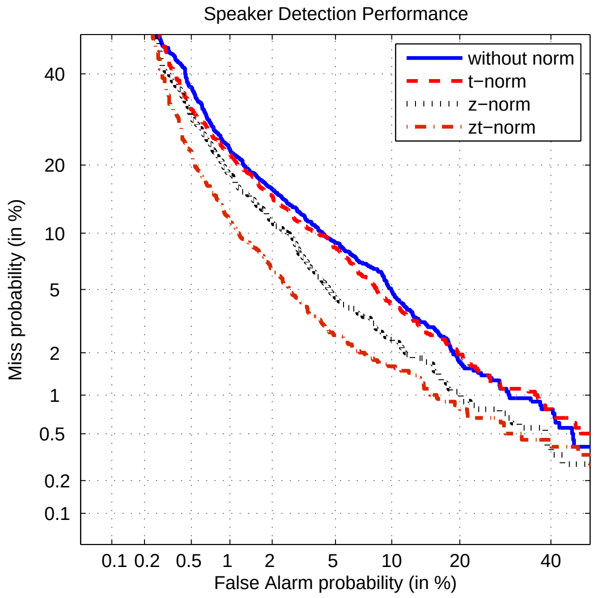

Figure 1: *DET curves for various types of score normalization using large numbers of imposters. Core condition of the NIST 2006 SRE, English language trials only.*

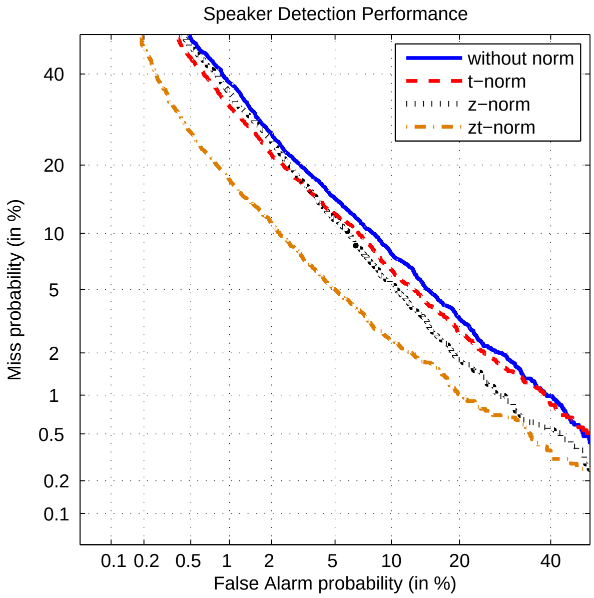

Figure 2: *DET curves for various types of score normalization using large numbers of imposters. Core condition of the NIST 2006 SRE, all trials.*

4. Score normalization

In setting thresholds for making verification decisions with factor analysis models, we have always found z-norm to be much more effective than t-norm and that zt-norm is (by far) the most effective score normalization procedure. This prompted us to explore the top-norm method proposed by Zigel and Wasserblat [11]. This is a modification of z-norm which consists in selecting, for each target speaker, an N-best list of imposters from a very large number of z-norm utterances (e.g. 1000). We also experimented with the analogous top-norm procedure in the case of t-norm, where for each test utterance, we select the top N t-norm speakers as imposters.

For the experiments in this section we used a factor analysis model with 300 speaker factors and 50 channel factors trained using the data sets described in Section 3.3 and the t-norm speakers and z-norm utterances described in Section 3.4; we used the core condition of the 2006 SRE for testing.

Our first concern was to find out what would happen if we implemented zt-norm in the usual way using the very large numbers of imposters described in Section 3.4. As a benchmark we used 100 z-norm utterances and 100 t-norm speakers and found that zt-norm gave an EER of 4.8% and a DCF of 0.025 on the English language trials in the core condition. Using 1000 z-norm utterances and several hundred t-norm speakers gave a substantial improvement, namely an EER of 3.5% and a DCF of 0.021. The DET curves corresponding to t-norm, z-norm and zt-norm are shown in Fig. 1.

We found a similar pattern when we replicated these experiments on all trials of the core condition rather than the English language subset. The benchmark results were 7.2% (EER) and 0.036 (DCF); using all of the imposters give an EER of 5.0% and a DCF of 0.027. The DET curves corresponding to t-norm, z-norm and zt-norm in this situation are shown in Fig. 2. The outstanding effectiveness of zt-norm compared with both z-norm and t-norm is apparent in both sets of DET curves.

The results we obtained when we replicated the top-norm experiment in [11] on the common subset (i.e. English language trials) of the core condition of the 2006 SRE are summarized in Table 1. The results in the z-norm column show that selecting the top 100 z-norm utterances for each target speaker is indeed more effective than using all of the z-norm utterances. However, the results in the zt-norm column show that restricting the number of z-norm speakers is harmful.

Table 1: *Top z-norm results on the English trials of core condition of the NIST 2006 SRE.*

	z-norm		zt-norm	
	EER	DCF	EER	DCF
top 100 z-norm	**4.6%**	**0.025**	3.8%	0.023
top 200 z-norm	4.6%	0.026	3.8%	0.022
top 300 z-norm	4.7%	0.027	3.8%	0.022
top 400 z-norm	4.7%	0.027	3.6%	0.022
top 500 z-norm	4.7%	0.027	3.6%	0.022
all z-norm	4.7%	0.027	**3.5%**	**0.021**

The question arises whether applying a similar top-norm selection to t-norm speakers could be beneficial. Because t-norm is computationally expensive, it is difficult to experiment with this idea on a large scale; the results in Table 2 indicate that small improvements might be achievable.

For completeness, we present the corresponding results on all trials of the core condition of the NIST 2006 SRE in Tables 3 and 4. Patterns similar to those in Table 1 can be observed in Table 3. The results in Table 4, like those in Table 2, again suggest that small improvements in performance might be obtained by applying top-norm to t-norm speakers but, for the experiments reported in the remainder of the paper, we used zt-norm with all of the imposters described in Section 3.4.

In conducting these experiments we observed that zt-norm

Table 2: *Top t-norm results on the English trials of core condition of the NIST 2006 SRE.*

	t-norm		zt-norm	
	EER	DCF	EER	DCF
top 100 t-norm	6.7%	0.030	3.5%	0.020
top 150 t-norm	6.6%	0.030	3.4%	0.020
top 200 t-norm	6.6%	0.032	3.6%	0.020
all t-norm	6.5%	0.031	3.5%	0.021

Table 3: *Top z-norm results on all trials of the core condition of the NIST 2006 SRE.*

	z-norm		zt-norm	
	EER	DCF	EER	DCF
top 100 z-norm	**7.3%**	**0.036**	5.4%	0.029
top 200 z-norm	7.3%	0.037	5.2%	0.029
top 300 z-norm	7.3%	0.037	5.2%	0.028
top 400 z-norm	7.3%	0.038	5.1%	0.028
top 500 z-norm	7.5%	0.039	5.1%	0.028
all z-norm	7.5%	0.043	**5.0%**	**0.027**

consistently resulted in large improvements compared with no score normalization. These improvements seem larger than those obtained by zt-norm in comparable systems so, after we had finished running all of our experiments, we returned to this question to see if we could get a better understanding of why zt-norm made such a big difference in our system.

One respect in which our speaker enrollment procedure differs from that of other systems is that it provides an estimate of the uncertainty in MAP estimation of a target speaker's supervector $\boldsymbol{s}$ which arises from the fact that the speaker's enrollment data is of limited duration. This uncertainty is expressed as a diagonal $CF \times CF$ covariance matrix which we denoted by $\mathrm{Cov}(\boldsymbol{s}, \boldsymbol{s})$ in [5]. The uncertainty is typically quite large (about 10% of the variance of the speaker population in a 300 speaker factor model and much larger in the case of no speaker factors) and it will only be zero if the amount of enrollment data is infinite.

We included the covariance matrix $\mathrm{Cov}(\boldsymbol{s}, \boldsymbol{s})$ in the function which evaluates the score of a verification trial (equation (19) in [5]). The effect of this is to decrease the score of a verification trial if the uncertainty in estimating the hypothesized speaker's supervector is large. (The idea is to penalize target speakers having small amounts of enrollment data.) In the light of the experiments conducted here, it occurred to us that zt-norm must actually be undoing the effect of including $\mathrm{Cov}(\boldsymbol{s}, \boldsymbol{s})$ in the likelihood evaluation because the purpose of z-norm is to fit the scores of imposter trials for *all* target speakers to a common bell curve (irrespective of the amounts of enrollment data available for the different target speakers). So we ran some experiments to see the effect of suppressing the contribution of these covariance matrices. For these experiments, we used the female portion of the NIST 2006 core condition (English language trials only). The results are summarized in Table 5. Note that, as we anticipated, uncertainty modeling has no effect whatever in the presence of zt-norm. However, it actually hurts in the absence of score normalization (line 3) so this seems to explain part of the large performance gains that we attributed to zt score normalization. Note that when uncertainty modeling is turned off, it is still the case that 1000 z-norm utterances give better results than a more 'reasonable' number, 222.

Table 4: *Top t-norm results on all trials of the core condition of the NIST 2006 SRE.*

	t-norm		zt-norm	
	EER	DCF	EER	DCF
top 100 t-norm	8.0%	0.039	4.8%	0.025
top 150 t-norm	7.9%	0.039	4.8%	0.026
top 200 t-norm	8.2%	0.042	4.8%	0.026
all t-norm	8.2%	0.041	5.0%	0.027

Table 5: *The effect of zt-score normalization with and without uncertainty modeling. Female portion of the NIST 2006 core condition, English language trials only.*

	With Uncertainty		Without	
	EER	DCF	EER	DCF
1000 z-norm utterances	4.1%	0.024	4.1%	0.024
222 z-norm utterances	4.5%	0.024	4.5%	0.024
no norm	7.7%	0.035	6.3%	0.032

We no longer use uncertainty modeling in our system. Although the results that we will report in the remainder of the paper were obtained with uncertainty modeling, there is no reason to believe that they were affected by this.

5. Varying the number of speaker factors in the core condition

We trained gender dependent factor analysis models, each having 50 channel factors but different numbers of speaker factors, on the data sets described in Section 3.3 and tested these models on the core condition of the NIST 2006 SRE. The results are summarized in Table 6 where it is apparent that the larger the number of speaker factors, the better the performance. (The expression "0 speaker factors" refers to the case where $\boldsymbol{v} = \mathbf{0}$ in (3); the expression $\boldsymbol{d} \neq \mathbf{0}$ indicates explicitly that the term $\boldsymbol{dz}$ was included in (3).)

Table 6: *Results obtained on the core condition of the NIST 2006 SRE with varying numbers of speaker factors.*

	All trials		English trials	
	EER	DCF	EER	DCF
$\boldsymbol{d} \neq \mathbf{0}$, 0 speaker factors	6.2%	0.030	4.6%	0.025
$\boldsymbol{d} \neq \mathbf{0}$, 100 speaker factors	5.3%	0.029	4.1%	0.024
$\boldsymbol{d} \neq \mathbf{0}$, 300 speaker factors	**5.0%**	**0.027**	**3.5%**	**0.021**

These results are not surprising; we found exactly the same pattern in our experiments on the core condition of the NIST 2005 SRE reported in [5]. A question which we did not attempt to explore in that article was whether performance would continue to improve as we increased the number of speaker factors beyond 300. The main obstacle here is the computational burden of training factor analysis models with very large numbers of speaker factors. This can be alleviated by modifying the training algorithm so as to suppress the term $\boldsymbol{dz}$ in (3) altogether and by using MAP estimates of the hidden variables rather than integrating over them. (The simplification here is analogous to the difference between the forward-backward algorithm and Viterbi decoding.) We rewrote our software accordingly but since the new version was designed to work with full covariance universal background models, and since we are using very large numbers of imposters in our experiments, it runs slowly at verification time and we have been restricted in the number of experiments we can perform with it. Thus we will only report results on the female portion of the test set. The results presented in Table 7 show that small improvements in the DCF can be obtained by increasing the number of speaker factors in this way, but for practical purposes, performance seems to saturate at about 300 speaker factors.

Table 7: *Results on the female portion of the core condition of the NIST 2006 SRE (English trials only) obtained with large numbers of speaker factors*

	English trials	
	EER	DCF
$\boldsymbol{d} \neq \mathbf{0}$, 300 speaker factors	4.1%	0.024
$\boldsymbol{d} = \mathbf{0}$, 500 speaker factors, full covariances	4.3%	0.023
$\boldsymbol{d} = \mathbf{0}$, 700 speaker factors, full covariances	4.2%	**0.022**

These results also suggest that only minor improvements (as measured by the DCF) can be obtained with full covariance GMM/UBM systems and that these improvements may be offset by minor degradations (as measured by the EER). This line of experimentation was partly motivated by the impressive results obtained with HLDA in [13]. If HLDA is implemented without dimensionality reduction (i.e. MLLT) then its role is to compensate for the inadequacies of the diagonal covariance matrix assumption in conventional Gaussian mixture modeling. But this assumption is probably no longer necessary: there is no shortage of speech data with which to train large full covariance UBM's for speaker recognition. In speech recognition, MLLT has already been superseded to some extent by full covariance HMM modeling. (IBM has developed speech recognizers with more than 100 K full covariance Gaussians [14].) Our preliminary results notwithstanding, it seems quite likely that full covariance GMM/UBM systems will eventually prove to be useful in speaker recognition. Another motivation for experimenting with full covariance UBM's was to see if we could find any useful information about speaker and/or channel effects in second order Baum-Welch statistics. However we had no success at all with that problem.

Finally we can compare our best results with the results submitted for the 2006 evaluation. (This is the "reference system" mentioned in Table 8.) The performance of the two systems is very similar but it is clear that excluding the 2005 SRE data from factor analysis training has cost us a few points in terms of the DCF on the English language trials of the core condition despite our improved score normalization and extended acoustic feature set.

Table 8: *Comparison of our best results on the core condition of the NIST 2006 SRE with the results submitted for the evaluation.*

	All trials		English trials	
	EER	DCF	EER	DCF
$\boldsymbol{d} \neq \mathbf{0}$, 300 speaker factors	5.0%	0.027	3.5%	0.021
$\boldsymbol{d} \neq \mathbf{0}$, Reference system	5.1%	0.027	3.3%	**0.017**

6. Varying the number of speaker factors in the extended data condition

We now report the results of replicating the experiments in Section 5 on the extended data condition of the 2006 NIST SRE.

Returning to the factor analysis models with 0, 100 and 300 speaker factors described in Section 5, Table 9 summarizes the results we obtained when we used these models in the extended data condition of the 2006 evaluation. Here we observe a different pattern from Table 6: the best performance is achieved with 100 rather than 300 speaker factors. Note also that, at least as measured by the EER, the performance with 100 speaker factors is much better than with no speaker factors.

Table 9: *Results obtained on the extended data condition of the NIST 2006 SRE with varying numbers of speaker factors.*

	All trials		English trials	
	EER	DCF	EER	DCF
$\boldsymbol{d} \neq \mathbf{0}$, 0 speaker factors	3.0%	0.014	2.7%	0.013
$\boldsymbol{d} \neq \mathbf{0}$, 100 speaker factors	**2.2%**	**0.011**	**1.9%**	**0.010**
$\boldsymbol{d} \neq \mathbf{0}$, 300 speaker factors	2.4%	0.012	2.1%	0.011

To interpret these results it is helpful to compare the the matrix traces $\mathrm{tr}(\boldsymbol{d}^2)$ and $\mathrm{tr}(\boldsymbol{v}\boldsymbol{v}^*)$ in the 100 and 300 speaker factor case. For the 100 speaker factor model (female gender) we found

$$\mathrm{tr}(\boldsymbol{d}^2) = 53.85, \quad \mathrm{tr}(\boldsymbol{v}\boldsymbol{v}^*) = 400.71 \tag{4}$$

and for the 300 speaker factor model

$$\mathrm{tr}(\boldsymbol{d}^2) = 10.77, \quad \mathrm{tr}(\boldsymbol{v}\boldsymbol{v}^*) = 465.52. \tag{5}$$

Thus, for the 300 speaker factor model, almost all of the speaker variability is confined to the speaker space (i.e. $\boldsymbol{m} +$ Range $\boldsymbol{v}\boldsymbol{v}^*$) and $\mathrm{tr}(\boldsymbol{d}^2)$ is much smaller than in the case of the 100 speaker factor model. (It is reasonable that most of the speaker variability should be accounted for by $\boldsymbol{v}$ in both cases since $\boldsymbol{v}$ has many more free parameters than $\boldsymbol{d}$.)

As was first pointed out in [15], even if $\boldsymbol{d}$ is small, it will always contribute something in estimating a speaker model provided that there is sufficient enrollment data for the speaker. For the 300 speaker factor model, $\boldsymbol{d}$ is so small that this effect is not manifest even if there are 15–20 minutes of enrollment data. On the other hand, for the 100 speaker factor model, both of the terms $\boldsymbol{d}\boldsymbol{z}$ and $\boldsymbol{v}\boldsymbol{y}$ in (3) are contributing in the extended data condition.

Again, the question arises whether increasing the dimensionality of the speaker space beyond 300 can improve performance. We obtained some results on the female subset of the extended data condition using the full covariance factor analysis models with 500 and 700 speaker factors described in Section 5. These results are summarized in Table 10; they show that performance has already saturated at 100 speaker factors.

Finally we can compare our best results with the results submitted for the 2006 evaluation. (This is the "reference system" mentioned in Table 11.) We see that the two sets of results are very similar on the English language trials of the extended data condition and that the results obtained with the rebuilt factor analysis system on all trials of the extended data condition are better than those of the reference system.

Table 10: *Results on the female portion of the extended data condition of the NIST 2006 SRE (English trials only) obtained with large numbers of speaker factors*

	English trials	
	EER	DCF
$\boldsymbol{d} \neq \mathbf{0}$, 100 speaker factors	1.9%	**0.010**
$\boldsymbol{d} \neq \mathbf{0}$, 300 speaker factors	2.1%	0.011
$\boldsymbol{d} = \mathbf{0}$, 500 speaker factors, full covariances	2.3%	0.014
$\boldsymbol{d} = \mathbf{0}$, 700 speaker factors, full covariances	2.1%	0.014

Table 11: *Comparison of our best results on the extended data condition of the NIST 2006 SRE with the results submitted for the evaluation.*

	All trials		English trials	
	EER	DCF	EER	DCF
$\boldsymbol{d} \neq \mathbf{0}$,100 speaker factors	**2.2%**	**0.011**	1.9%	0.010
$\boldsymbol{d} \neq \mathbf{0}$, Reference system	2.4%	0.015	1.7%	0.009

7. Conclusions

A factor analysis system with several hundred speaker factors that has been trained with data consisting of multiple recordings of, say, 1000 speakers is capable of memorizing the characteristics of individual speakers in the training set so that, if the system is tested on the training speakers, the results can be quite misleading. We first observed this phenomenon in [6].

Some of the data from the 2005 NIST SRE was recycled for the extended data task of the 2006 evaluation. Since we used this data for factor analysis training, there was some doubt about the validity of the results we submitted.

We therefore decided to exclude the 2005 data from factor analysis training in order to experiment properly with the 2006 extended data test set. The restricted training set consists principally of Switchboard rather than Mixer data so we did not expect to obtain particularly good results. However we found (as in [5]) that some simple experiments in score normalization gave dramatic improvements in performance. Thus by using a very large number of z-norm speakers for zt-norm we were able to obtain results comparable to those we submitted at the time of the 2006 evaluation, both on the core condition and on the extended data condition.

A general trend which is apparent in the results presented in this paper is that we obtained substantially better performance on English language trials than on all trials. We attribute this to the fact that our factor analysis training set contains very little non-English speech (more than 80% of the data comes from the Switchboard corpora). Another clear pattern is that performance on female speakers is much poorer than on males. We had hoped that increasing the number of cepstral features from 12 to 19 would narrow the gender gap.

Our experience with the core condition of the previous NIST SRE data sets has been that the eigenvoice component $\boldsymbol{vy}$ in the speaker variability model (3) is much more useful than the classical MAP component $\boldsymbol{dz}$ for speaker modeling. The results presented in Section 5 provide more confirmation of this.

On the other hand the NIST extended data task is interesting from the point of view of factor analysis modeling because it facilitates experimentation with this question in situations where very large amounts (15–20 minutes) of enrollment data are available for a target speaker. We therefore experimented with a range of speaker factor configurations and found that the best performance could be achieved with a limited number of speaker factors (100 rather than 300 or more as in the core condition) and that both terms $\boldsymbol{vy}$ and $\boldsymbol{dz}$ contribute to speaker modeling in this situation.

The reason for the fact that the term $\boldsymbol{dz}$ in (3) is of limited use is that almost all of the inter-speaker variability in a factor analysis training set can be accounted for with a sufficiently large number of speaker factors. This led us to revisit the question of how best to estimate $\boldsymbol{d}$ in [16]. In that paper, we propose a new estimation procedure for $\boldsymbol{d}$ which yields 10–15% reductions in error rates on the core condition of the NIST 2006 SRE as well as on the extended data condition.

8. References

[1] J.-L. Gauvain and C.-H. Lee, “Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.

[2] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[3] P. Kenny, “Joint factor analysis of speaker and session variability: Theory and algorithms, Tech. Report CRIM-06/08-13,” 2005.

[4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Speaker and session variability in GMM-based speaker verification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, May 2007.

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, May 2007.

[6] P. Kenny and P. Dumouchel, “Disentangling speaker and channel effects in speaker verification,” in *Proc. ICASSP*, Montreal, Canada, May 2004.

[7] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, May 2005.

[8] N. Dehak, P. Kenny, and P. Dumouchel, “Modeling prosodic features with joint factor analysis for speaker verification,” *IEEE Trans. Audio, Speech, Lang. Process.*, Sept. 2007.

[9] “The NIST year 2006 speaker recognition evaluation plan,” 2006.

[10] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *Proc. Speaker Odyssey*, Crete, Greece, June 2001, pp. 213–218.

[11] Yaniv Zigel and Moshe Wasserblat, “How to deal with multiple-targets in speaker identification systems?,” in *Proc. IEEE Odyssey 2006*, San Juan, Puerto Rico, June 2006.

[12] P. Kenny and S.-C. Yin, “The CRIM systems for the 2006 NIST speaker recognition evaluation,” in *The 2006 NIST Speaker Recognition Workshop*, San Juan, Puerto Rico, June 2006.

[13] Lukas Burget, Pavel Matejka, Ondrej Glembek, Petr Schwarz, and Jan Cernocky, "Analysis of feature extraction and channel compensation in GMM speaker recognition system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 2007, no. 12, pp. 7, 2007.

[14] Stanley F. Chen, Brian Kingsbury, Lidia Mangu, Daniel Povey, George Saon, Hagen Soltau, and Geoffrey Zweig, "Advances in speech transcription at IBM under the DARPA EARS program," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 5, pp. 1596–1608, 2006.

[15] Simon Lucey and Tsuhan Chen, "Improved speaker verification through probabilistic subspace adaptation," in *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003, pp. 2021–2024.

[16] P. Kenny, N. Dehak, V. Gupta, and P. Dumouchel, "A new training regimen for factor analysis of speaker variability," in *Proc. ICASSP 2008*, Las Vegas, Nevada (submitted), Mar. 2008.