

A NEW TRAINING REGIMEN FOR FACTOR ANALYSIS OF SPEAKER VARIABILITY

Patrick Kenny, Najim Dehak, Vishwa Gupta, Pierre Ouellet and Pierre Dumouchel

Centre de recherche informatique de Montréal (CRIM)

{Patrick.Kenny, Najim.Dehak, Vishwa.Gupta, Pierre.Ouellet, Pierre.Dumouchel}@crim.ca

ABSTRACT

We propose a new approach to the problem of estimating speaker factor loading matrices which enhances the effectiveness of the classical MAP component of factor analysis. Our past experience led us to believe that this component was only useful in extended data tasks where 15–20 mins of speech data is available to enroll target speakers. When the proposed estimation technique was tested on the NIST 2006 speaker recognition evaluation data, we found that the classical MAP component was responsible for 10–15% reductions in error rates on the core condition (as measured both by equal error rates and the NIST detection cost function). Similar improvements were obtained on the extended data task.

Index terms speaker recognition

1. INTRODUCTION

Factor analysis is a model of speaker and session variability in Gaussian mixture models (GMM's) which is widely used in text-independent speaker recognition. This paper is concerned with the speaker variability component of our version of factor analysis. The role of this component is to provide a prior distribution for *maximum a priori* (MAP) estimation of speaker-dependent GMM's at enrollment time.

Let F be the dimension of the acoustic feature vectors and let C be the number of mixture components in the universal background model (UBM); we index these mixture components by $c = 1, \dots, C$. We use the term speaker supervector to refer to the CF dimensional vector obtained by concatenating the F dimensional mean vectors in a speaker-dependent GMM.

If s is the supervector for a randomly chosen speaker then we assume that s is distributed according to

$$s = m + vy + dz \quad (1)$$

where m is a speaker-independent supervector, d is a diagonal matrix, v is a rectangular matrix of low rank and y and z are independent random vectors having standard normal distributions. (This assumption is equivalent to saying that s is normally distributed with mean m and covariance matrix $d^2 + vv^*$.) Also we associate a diagonal covariance matrix Σ_c with each mixture component c whose role is to model the variability in the acoustic observation vectors which is not captured by the speaker variability model (1). We denote by Σ the $CF \times CF$ supercovariance matrix whose diagonal is the concatenation of these covariance matrices.

The components of y are the speaker factors. We will write $y(s)$ in place of y in situations where it is necessary to refer to the speaker factors associated with a particular speaker s . The speaker variability model (1) reduces to classical MAP in the case $v = 0$

and to eigenvoice MAP [1] in the case $d = 0$. Given the hyperparameters m, v, d, Σ together with a channel factor loading matrix u , the calculation of the MAP estimate of s can be carried out as in Section III of [2].

Generally speaking, our experience has been that the term vy in (1) is much more useful than the term dz . In a typical factor analysis training scenario with, say, 1000 training speakers and 300 speaker factors, almost all of the speaker variability in the training set can be well accounted for by v alone (v has 300 times as many free parameters as d). Thus, if care is not taken, it can happen that d ends up playing no useful role. The principal conclusion that we came to in [3] was that the term dz is really only useful in extended data tasks where 15–20 minutes of enrollment data are available for each target speaker. This led us to re-examine the algorithms that we used to estimate v and d . In this paper we propose a new training regimen for these hyperparameters which enabled us to obtain 10–15% reductions in error rates on the core condition of the NIST 2006 speaker recognition evaluation data as well as on the extended data task.

2. ESTIMATING THE HYPERPARAMETERS

The UBM supervector can serve as an estimate of m and the UBM covariance matrices are good first approximations to the residual covariance matrices Σ_c ($c = 1, \dots, C$). The problem of estimating v in the case where $d = 0$ was addressed in [1] and a very similar approach can be adopted to estimating d in the case where $v = 0$. We first summarize the estimation procedures for these two special cases and then explain how they can be combined to tackle the general case.

2.1. Baum-Welch statistics

Given a speaker s and acoustic feature vectors Y_1, Y_2, \dots , for each mixture component c we define the Baum-Welch statistics in the usual way:

$$\begin{aligned} N_c(s) &= \sum_t \gamma_t(c) \\ F_c(s) &= \sum_t \gamma_t(c) Y_t \\ S_c(s) &= \text{diag} \left(\sum_t \gamma_t(c) Y_t Y_t^* \right) \end{aligned}$$

where, for each time t , $\gamma_t(c)$ is the posterior probability of the event that the feature vector Y_t is accounted for by the given mixture component. We calculate these posteriors using the UBM (other possibilities could be explored).

We denote the centralized first and second order Baum-Welch statistics by $\tilde{F}_c(s)$ and $\tilde{S}_c(s)$:

$$\begin{aligned}\tilde{F}_c(s) &= \sum_t \gamma_t(c)(Y_t - m_c) \\ \tilde{S}_c(s) &= \text{diag} \left(\sum_t \gamma_t(c)(Y_t - m_c)(Y_t - m_c)^* \right)\end{aligned}$$

where m_c be the subvector of \mathbf{m} corresponding to the mixture component c . In other words,

$$\begin{aligned}\tilde{F}_c(s) &= F_c(s) - N_c(s)m_c \\ \tilde{S}_c(s) &= S_c(s) \\ &\quad - \text{diag} (F_c(s)m_c^* + m_c F_c(s)^* - N_c(s)m_c m_c^*).\end{aligned}$$

Let $N(s)$ be the $CF \times CF$ diagonal matrix whose diagonal blocks are $N_c(s)I$ ($c = 1, \dots, C$). Let $\tilde{\mathbf{F}}(s)$ be the $CF \times 1$ supervector obtained by concatenating $\tilde{F}_c(s)$ ($c = 1, \dots, C$). Let $\tilde{\mathbf{S}}(s)$ be the $CF \times CF$ diagonal matrix whose diagonal blocks are $\tilde{S}_c(s)$ ($c = 1, \dots, C$).

2.2. Training an eigenvoice model

In this section we consider the problem of estimating \mathbf{m}, \mathbf{v} and Σ under the assumption that $\mathbf{d} = \mathbf{0}$. We assume that initial estimates of the hyperparameters are given. (Random initialization of \mathbf{v} works fine in practice.)

2.2.1. The posterior distribution of the hidden variables

For each speaker s , set $\mathbf{l}(s) = \mathbf{I} + \mathbf{v}^* \Sigma^{-1} N(s) \mathbf{v}$. Then the posterior distribution of $\mathbf{y}(s)$ conditioned on the acoustic observations of the speaker is Gaussian with mean $\mathbf{l}^{-1}(s) \mathbf{v}^* \Sigma^{-1} \tilde{\mathbf{F}}(s)$ and covariance matrix $\mathbf{l}^{-1}(s)$. (See [1], Proposition 1.)

We will use the notation $E[\cdot]$ to indicate posterior expectations; thus $E[\mathbf{y}(s)]$ denotes the posterior mean of $\mathbf{y}(s)$ and $E[\mathbf{y}(s)\mathbf{y}^*(s)]$ the posterior correlation matrix.

2.2.2. Maximum likelihood re-estimation

This entails accumulating the following statistics over the training set where the posterior expectations are calculated using initial estimates of $\mathbf{m}, \mathbf{d}, \Sigma$ and s ranges over the training speakers:

$$\begin{aligned}N_c &= \sum_s N_c(s) \quad (c = 1, \dots, C) \\ \mathfrak{A}_c &= \sum_s N_c(s) E[\mathbf{y}(s)\mathbf{y}^*(s)] \quad (c = 1, \dots, C) \\ \mathfrak{C} &= \sum_s \tilde{\mathbf{F}}(s) E[\mathbf{y}^*(s)] \\ N &= \sum_s N(s).\end{aligned}$$

For each mixture component $c = 1, \dots, C$ and for each $f = 1, \dots, F$, set $i = (c-1)F + f$ let v_i denote the i th row of \mathbf{v} and \mathfrak{C}_i the i th row of \mathfrak{C} . Then \mathbf{v} is updated by solving the equations

$$v_i \mathfrak{A}_c = \mathfrak{C}_i \quad (i = 1, \dots, CF).$$

The update formula for Σ is

$$\Sigma = N^{-1} \left(\sum_s \tilde{\mathbf{S}}(s) - \text{diag}(\mathfrak{C}\mathbf{v}^*) \right).$$

(See [1], Proposition 3.)

2.2.3. Minimum divergence re-estimation

Given initial estimates \mathbf{m}_0 and \mathbf{v}_0 , the update formulas for \mathbf{m} and \mathbf{v} are

$$\begin{aligned}\mathbf{m} &= \mathbf{m}_0 + \mathbf{v}_0 \mu_{\mathbf{y}} \\ \mathbf{v} &= \mathbf{v}_0 \mathbf{T}_{\mathbf{y}\mathbf{y}}^*\end{aligned}$$

Here

$$\mu_{\mathbf{y}} = \frac{1}{S} \sum_s E[\mathbf{y}(s)],$$

$\mathbf{T}_{\mathbf{y}\mathbf{y}}$ is an upper triangular matrix such that

$$\mathbf{T}_{\mathbf{y}\mathbf{y}}^* \mathbf{T}_{\mathbf{y}\mathbf{y}} = \frac{1}{S} \sum_s E[\mathbf{y}(s)\mathbf{y}^*(s)] - \mu_{\mathbf{y}} \mu_{\mathbf{y}}^*$$

(i.e. Cholesky decomposition), S is the number of training speakers, and the sums extend over all speakers in the training set. (See [2], Theorem 7.) The role of this type of estimation is to get good estimates of the eigenvalues corresponding to the eigenvoices.

2.3. Training a diagonal model

An analogous development can be used to estimate \mathbf{m}, \mathbf{d} and Σ if \mathbf{v} is constrained to be $\mathbf{0}$.

2.3.1. The posterior distribution of the hidden variables

For each speaker s , set $\mathbf{l}(s) = \mathbf{I} + \mathbf{d}^2 \Sigma^{-1} N(s)$. Then the posterior distribution of $\mathbf{z}(s)$ conditioned on the acoustic observations of the speaker is Gaussian with mean $\mathbf{l}^{-1}(s) \mathbf{d} \Sigma^{-1} \tilde{\mathbf{F}}(s)$ and covariance matrix $\mathbf{l}^{-1}(s)$.

Again, we will use the notation $E[\cdot]$ to indicate posterior expectations; thus $E[\mathbf{z}(s)]$ denotes the posterior mean of $\mathbf{z}(s)$ and $E[\mathbf{z}(s)\mathbf{z}^*(s)]$ the posterior correlation matrix.

It is straightforward to verify that, in the special case where \mathbf{d} is assumed to satisfy

$$\mathbf{d}^2 = \frac{1}{r} \Sigma,$$

this posterior calculation leads to the standard relevance MAP estimation formulas for speaker supervectors (r is the relevance factor). The following two sections summarize data-driven procedures for estimating \mathbf{m}, \mathbf{d} and Σ which do not depend on the relevance MAP assumption. It can be shown that when these update formulas are applied iteratively, the values of a likelihood function analogous to that given in Proposition 2 of [1] increase on successive iterations.

2.3.2. Maximum likelihood re-estimation

This entails accumulating the following statistics over the training set where the posterior expectations are calculated using initial estimates of \mathbf{m} , \mathbf{d} , Σ and s ranges over the training speakers:

$$\begin{aligned} N_c &= \sum_s N_c(s) \quad (c = 1, \dots, C) \\ \mathbf{a} &= \sum_s \text{diag}(\mathbf{N}(s)E[\mathbf{z}(s)\mathbf{z}^*(s)]) \\ \mathbf{b} &= \sum_s \text{diag}(\tilde{\mathbf{F}}(s)E[\mathbf{z}^*(s)]) \\ \mathbf{N} &= \sum_s \mathbf{N}(s). \end{aligned}$$

For $i = 1, \dots, CF$ let d_i the i th entry of \mathbf{d} and similarly for a_i and b_i . Then \mathbf{d} is updated by solving the equation

$$d_i a_i = b_i$$

for each i . The update formula for Σ is

$$\Sigma = \mathbf{N}^{-1} \left(\sum_s \tilde{\mathbf{S}}(s) - \text{diag}(\mathbf{b}\mathbf{d}) \right).$$

2.3.3. Minimum divergence re-estimation

Given initial estimates \mathbf{m}_0 and \mathbf{d}_0 , the update formulas for \mathbf{m} and \mathbf{d} are

$$\begin{aligned} \mathbf{m} &= \mathbf{m}_0 + \mathbf{d}_0 \boldsymbol{\mu}_z \\ \mathbf{d} &= \mathbf{d}_0 \mathbf{T}_{zz} \end{aligned}$$

where

$$\boldsymbol{\mu}_z = \frac{1}{S} \sum_s E[\mathbf{z}(s)],$$

\mathbf{T}_{zz} is a diagonal matrix such that

$$\mathbf{T}_{zz}^2 = \text{diag} \left(\frac{1}{S} \sum_s E[\mathbf{z}(s)\mathbf{z}^*(s)] - \boldsymbol{\mu}_z \boldsymbol{\mu}_z^* \right),$$

S is the number of training speakers, and the sums extend over all speakers in the training set.

We will need a variant of this update procedure which applies to the case where \mathbf{m} is forced to be $\mathbf{0}$. In this case \mathbf{d} is estimated from \mathbf{d}_0 by taking \mathbf{T}_{zz} to be such that

$$\mathbf{T}_{zz}^2 = \text{diag} \left(\frac{1}{S} \sum_s E[\mathbf{z}(s)\mathbf{z}^*(s)] \right).$$

2.4. Joint estimation of \mathbf{v} and \mathbf{d}

There is no difficulty in principle in extending the maximum likelihood and minimum divergence training procedures to handle a general factor analysis model in which both \mathbf{v} and \mathbf{d} are non-zero (Theorems 4 and 7 in [2]). We used this type of joint estimation in all of our previous work in factor analysis and to produce benchmarks for the experiments that we will report in this paper.

However joint estimation of \mathbf{v} and \mathbf{d} is computationally demanding because, in a general factor analysis model, all of the

hidden variables become correlated with each other in the posterior distributions. Our experience has been that, given the Baum-Welch statistics, training a diagonal model runs very quickly and training a pure eigenvoice model can be made to run quickly (at the cost of some memory overhead) by suitably organizing the computation of the matrices $\mathbf{l}(s)$ in Section 2.2.1. Unfortunately no such computational short cuts seem to be possible in the general case. Furthermore, even if the eigenvoice component \mathbf{v} is carefully initialized, many iterations of joint estimation seem to be needed to estimate \mathbf{d} properly and, because the contribution of \mathbf{d} to the likelihood of the training data is minor compared with the contribution of \mathbf{v} , it is difficult to judge when the training algorithm has effectively converged.

2.5. Decoupled estimation of \mathbf{v} and \mathbf{d}

These considerations together with the fact that \mathbf{d} has far fewer free parameters than \mathbf{v} led us to explore an alternative training regimen where we divide the training speakers into two disjoint sets. We use the larger of the two sets to estimate \mathbf{m} and \mathbf{v} and the smaller to estimate \mathbf{d} and Σ .

Specifically, we first fit a pure eigenvoice model to the larger training set using the procedures described in Sections 2.2.2 and 2.2.3. Then, for each speaker s in the residual training set, we calculate the MAP estimate of $\mathbf{y}(s)$, namely $E[\mathbf{y}(s)]$, as in Section 2.2.1. This gives us a preliminary estimate of the speaker's supervector s , namely

$$s = \mathbf{m} + \mathbf{v}E[\mathbf{y}(s)].$$

We centralize the speaker's Baum-Welch statistics by subtracting the speaker's supervector (that is we apply the formulas in Section 2.1 with \mathbf{m} replaced by s). Finally we use these centralized statistics together with the procedures described in Sections 2.3.2 and 2.3.3 to estimate a pure diagonal model with $\mathbf{m} = \mathbf{0}$. This gives us estimates of \mathbf{d} and Σ .

Since this training algorithm uses only the diagonal and eigenvoice estimation procedures, it converges rapidly.

3. EXPERIMENTS

3.1. Enrollment and test data

We used the core condition and the 8 conversation training condition (also known as the extended data condition) of the NIST 2006 speaker recognition evaluation (SRE) for testing. Although we will report results on male speakers as well as female, we used only the female trials for our experiments.

3.2. Feature Extraction

We extracted 19 cepstral coefficients together with a log energy feature using a 25 ms Hamming window and a 10 ms frame advance. These were subjected to feature warping 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 UBM's having 1024 Gaussians and gender dependent factor analysis models having 0,100 and 300 speaker factors. 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, Parts 1 and 2; the NIST 2003 Language recognition evaluation data set; and the NIST 2004 SRE enrollment and test data.

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 decoupled estimation of v and d , we estimated v on the Switchboard data and d on the 2004 SRE data.

3.4. Imposters

The verification decision scores obtained with the factor analysis models were normalized using zt -norm. As in [3], 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 imposters were chosen at random from the factor analysis training data.

3.5. Results

The results of our experiments on the female portion of the common subset of core condition of the NIST 2006 SRE are summarized in Table 1. (EER refers to the equal error rate and DCF to the minimum value of the NIST detection cost function.) There are some blank entries in the table because decoupled estimation only applies only in the case where both v and d are non-zero. The best result is obtained with 300 speaker factors and decoupled

Table 1. Results obtained on the core condition of the NIST 2006 SRE (female speakers, English language trials)

	Joint Estimation		Decoupled	
	EER	DCF	EER	DCF
100 speaker factors, $d \neq 0$	4.4%	0.027	3.9%	0.022
300 speaker factors, $d \neq 0$	4.1%	0.024	3.6%	0.021
300 speaker factors, $d = 0$	3.9%	0.024	–	–
0 speaker factors, $d \neq 0$	5.2%	0.027	–	–

estimation. There is an anomaly in the joint estimation column: In the 300 speaker factor case we obtained a better EER by setting $d = 0$ than by joint estimation of v and d . We attribute this to the convergence issue mentioned in Section 2.4. Table 2 gives the corresponding results on all trials of the female portion of the core condition. Again 300 speaker factors with decoupled estimation gives the best results.

Table 2. Results obtained on the core condition of the NIST 2006 SRE (female speakers, all trials)

	Joint Estimation		Decoupled	
	EER	DCF	EER	DCF
100 speaker factors, $d \neq 0$	5.9%	0.032	4.9%	0.027
300 speaker factors, $d \neq 0$	5.6%	0.030	4.6%	0.025
300 speaker factors, $d = 0$	5.2%	0.028	–	–
0 speaker factors, $d \neq 0$	7.2%	0.034	–	–

We replicated these experiments on the female trials of the extended data condition. The results are summarized in Tables 3 and 4. Patterns similar to those in Tables 1 and 2 are evident.

For completeness, we report results on male speakers and on all speakers in Table 5.

Table 3. Results obtained on the extended data condition of the NIST 2006 SRE (female speakers, English language trials)

	Joint Estimation		Decoupled	
	EER	DCF	EER	DCF
100 speaker factors, $d \neq 0$	2.2%	0.012	2.1%	0.011
300 speaker factors, $d \neq 0$	2.1%	0.014	1.9%	0.011
300 speaker factors, $d = 0$	2.1%	0.014	–	–
0 speaker factors, $d \neq 0$	3.1%	0.017	–	–

Table 4. Results obtained on the extended data condition of the NIST 2006 SRE (female speakers, all trials)

	Joint Estimation		Decoupled	
	EER	DCF	EER	DCF
100 speaker factors, $d \neq 0$	2.5%	0.012	2.3%	0.014
300 speaker factors, $d \neq 0$	2.7%	0.014	2.3%	0.012
300 speaker factors, $d = 0$	2.7%	0.015	–	–
0 speaker factors, $d \neq 0$	3.6%	0.016	–	–

3.6. Conclusion

We have shown that decoupled estimation of v and d leads to 10–15% reductions in error rates compared with joint estimation on both the core condition and the extended data condition of the NIST 2006 SRE. The improvements that we observed were particularly marked on non-English trials. It seems that eigenvoice modeling successfully captures variability among English speakers and d is especially helpful with non-English speakers. Contrary to our conclusion in [3], the term dz in (1) can play a useful role in restricted data tasks after all.

4. REFERENCES

- [1] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, May 2005.
- [2] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms, Tech. Report CRIM-06/08-13," 2005. [Online]. Available: <http://www.crim.ca/perso/patrick.kenny>
- [3] P. Kenny, N. Dehak, R. Dehak, V. Gupta, and P. Dumouchel, "The role of speaker factors in the NIST extended data task," *Proc. IEEE Odyssey Workshop*, Stellenbosch, South Africa, Jan. 2008.

Table 5. Results obtained on the core condition and the extended data condition of the NIST 2006 SRE for male speakers and for all speakers using gender-dependent factor analysis models with 300 speaker factors (decoupled estimation)

	Male speakers		All speakers	
	EER	DCF	EER	DCF
Core condition, English	2.1%	0.013	3.1%	0.018
Core condition, all trials	4.2%	0.020	4.3%	0.023
Extended data, English	1.4%	0.006	1.8%	0.009
Extended data, all trials	1.7%	0.008	2.2%	0.011