# Factor Analysis with Sampling Methods for Text Dependent Speaker Recognition

*Antonio Miguel*[1], *Jesús Villalba*[1], *Alfonso Ortega*[1],
*Eduardo Lleida*[1], *Carlos Vaquero*[2]

[1] Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
[2] Agnitio S.L., Madrid, Spain

{amiguel, villalba, ortega, lleida}@unizar.es, cvaquero@agnitio-corp.com

10.21437/Interspeech.2014-332

## Abstract

Factor analysis is a method for embedding high dimensional data into a lower dimensional factor space. When data are multimodal we use mixtures of factor analyzers (MFA), which assume statistically independent samples. In speaker recognition, samples are not independent because they depend on the speaker in the utterance. In joint factor analysis and i-vectors, the MFA latent factors are tied at different levels. For example, they can be tied for a segment to extract utterance level information. Tied MFA approaches usually present the drawback that computing the exact posterior of the hidden variables (component responsibilities and latent factors) is unfeasible. For JFA, the preferred approximation consists in computing the responsibilities given a speaker independent GMM and they are fixed during the rest of the process. That implies that the estimated responsibilities for a given sample are independent of the rest of the samples of the utterance not taking into account the shared speaker and channel. We present a novel approximation to jointly estimate responsibilities and latent factors based on sampling the latent factor space. This model differs from previous ones in the hidden variables and parameter estimation; and likelihood evaluation. This approach was tested on the RSR2015 database for text-dependent speaker recognition

**Index Terms**: speaker recognition, factor analysis, sampling methods

## 1. Introduction

Many well known models for multimodal data or dimensionality reduction are formulated for statistically independent samples. When models like mixtures of Gaussians or mixtures of factor analyzers (MFA) [1] are applied to tasks such as speaker recognition [2, 3], they may ignore useful information at the sentence level, since they lack of a mechanism to model sequence information. In joint factor analysis (JFA) [4, 5], the MFA latent factors are tied at different levels: utterance or speaker. Although a very succinct Bayesian network is able to express tied MFA approaches, the model complexity is extremely high. Therefore, the exact calculation of the hidden variables posteriors and the exact evaluation of the likelihood becomes intractable.

We present a novel approximation to jointly estimate responsibilities and latent factors based on sampling the latent factor space. The new approach differs from previous joint factor analysis techniques in several aspects: how model parameters are estimated; how the inference of the posterior probabilities are calculated and how the likelihood is evaluated.

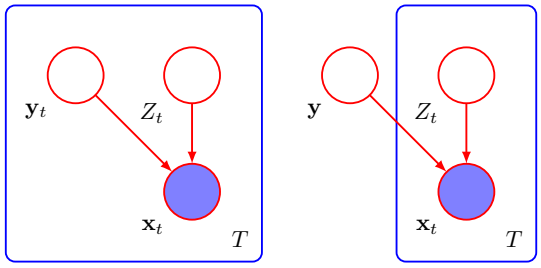
There are previous works that focus on embedding model parameters into a smaller dimensional subspace. Eigenvoices technique describes speaker dependent model parameters as linear combinations of the principal directions of variability of a population. It has been applied to fast adaptation of speaker models in speaker and speech recognition tasks [6, 7]. JFA proposed by Patrick Kenny [8, 5], is a more powerful tool, since it takes into account two levels of sequence variability: speaker and channel. JFA is a crucial part of the state-of-the-art speaker recognition systems [4, 5]. Low dimensional embedding factors at sentence level —known as i-vectors or total variability factors [9, 10, 11, 12]— can also be used as features for other systems where the different levels of variability are modeled, as it is done in face recognition, by using probabilistic linear discriminant analysis (PLDA) [13]. They are also used as features in many applications involving audio sequences for classification, detection or diarization systems [14, 15, 16, 17, 18]. The success of i-vectors evidences the importance of sequence level factors, since it has been shown by those works that they suffice to explain the underlying common information of a sequence, which can be kept to model classes or discarded to compensate channel effects.

Our approach differs from previous works in the treatment of the posteriors of the latent variables. For JFA, a speaker independent Gaussian mixture model (GMM) is used to calculate the responsibilities, which are thereafter fixed during the rest of the training process. If responsibilities are sample independent, then they do not take into account the shared speaker and channel latent information. In this work we study an approximation of not doing that assumption at different stages of the model training and evaluation, since the responsibilities are not independent if the common hidden factors are unknown. We will refer to this model as Tied Mixture of Factor Analyzers (TMFA).

This paper is organized as follows. In Section 2 MFAs are revised. Section 3 presents the TMFA model. Section 4 presents an experimental study to show the accuracy of the models in text dependent speaker recognition. Conclusions are presented in Section 5.

## 2. Mixtures of Factor Analyzers

In Figure 1a, we show the generative model of the MFA, where the observed data are a sequence $\bar{\mathbf{X}}$ of random vectors $\mathbf{x}_t \in \mathbb{R}^D$, which depend on the hidden factors $\mathbf{y}_t \in \mathbb{R}^R$, with $R < D$. The discrete hidden variables $Z_t \in \{1, \ldots, C\}$, with $C$ the number of components, provide the mechanism to generate/explain different modes in the data. The variables in the




(a) (b)

Figure 1: *Bayesian networks corresponding to sequence models: a) Mixture of Factor Analyzers (MFA), b) Tied Mixture of Factor Analyzers (TMFA)*

model are assumed to follow

$$\begin{aligned} \mathbf{y}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) && (1)\\ \mathbf{x}_t|\mathbf{y}_t, Z_t = c &\sim \mathcal{N}(\boldsymbol{\mu}_c + \mathbf{W}_c\mathbf{y}_t, \boldsymbol{\Psi}_c) && (2)\\ Z_t &\sim Cat(C, \boldsymbol{\pi}), && (3) \end{aligned}$$

where the distributions $\mathbf{x}|\mathbf{y}_t, Z_t = c$ are component specific; $\boldsymbol{\mu}_c$ are the mean vectors, $\boldsymbol{\Psi}_c$ are the covariance matrices of the intrinsic variability and $\mathbf{W}_c \in \mathbb{R}^{D\times R}$ are the factor loading matrices which span the factor subspace within the feature vector space; $Z_t$ is a discrete variable which follows a categorical or discrete distribution

$$p(Z_t = c) = \prod_{i=1}^{C} \pi_i^{\delta_{c,i}}, \tag{4}$$

where $\delta$ is the Kronecker's delta.

Due to the implicit frame independence in the MFA model, the likelihood of a sequence is the product $p(\bar{\mathbf{X}}) = \prod_t p(\mathbf{x}_t)$. To calculate the likelihood for a given sample $\mathbf{x}_t$, we integrate out the hidden variables $\mathbf{y}$ and $Z_t$ to obtain

$$\begin{aligned} p(\mathbf{x}_t) &= \sum_{c=1}^{C} \int_{\mathbf{y}} p(\mathbf{x}_t, \mathbf{y}, Z_t = c)\, \mathrm{d}\mathbf{y} \\ &= \sum_{c} \int_{\mathbf{y}} p(\mathbf{x}_t|\mathbf{y}, Z_t = c) p(Z_t = c) p(\mathbf{y})\, \mathrm{d}\mathbf{y} \\ &= \sum_{c} \pi_c \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \end{aligned} \tag{5}$$

where $\boldsymbol{\Sigma}_c = \mathbf{W}_c\mathbf{W}_c^\intercal + \boldsymbol{\Psi}_c$. We can see that the model is equivalent to a mixture of Gaussians with constrained covariance matrices that lay between a diagonal and a full covariance matrix.

## 3. Tied Mixtures of Factor Analyzers

Figure 1b shows the generative model of a TMFA. The main difference is that the latent factor of a standard MFA is sample dependent, while in the tied model it is shared by all the frames of the sequence. This apparently simple change in the directed graph configuration makes the inference of the hidden variables a very complex process due the large number of nodes having the same latent variable as parent. Therefore, we need to take approximations for practical applications.

In the case of the TMFA model, the hidden variable $\mathbf{y}$ plays a fundamental role in the model, since it is linked to all the observed samples in the sequence. We assume the following distributions for the model variables

$$\begin{aligned} \mathbf{y} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) && (6)\\ \mathbf{x}_t|\mathbf{y}, Z_t = c &\sim \mathcal{N}(\boldsymbol{\mu}_c + \mathbf{W}_c\mathbf{y}, \boldsymbol{\Psi}_c) && (7)\\ Z_t &\sim Cat(C, \boldsymbol{\pi}), && (8) \end{aligned}$$

where we have dropped the dependence of the hidden variable $\mathbf{y}$ with $t$. In this work, we will assume $\boldsymbol{\Psi}_c$ covariance matrices to be diagonal and we will refer to them as the internal or within-segment covariances, since they explain the speaker dependent intrinsic variability, as we can see in (7). Unlike usual implementations of JFA, in this work they are not assumed to be equal to the covariances of the speaker independent model.

### 3.1. Likelihood computation

The extra complexity of TMFA arises when we integrate out the hidden variables, since the component hidden variables $Z_t$ are not independent if the hidden factors are unknown [19]. Therefore, we have to use the joint distribution $p(\mathbf{x}_1, \ldots, \mathbf{x}_T)$. We denote $\bar{\mathbf{X}} = (\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_T^\intercal)^\intercal$ to the concatenation of all the frames in the sequence, with $\bar{\mathbf{X}} \in \mathbb{R}^{DT}$ being a column vector containing the whole utterance; we will refer to these concatenated vectors as super vectors. Integrating out the hidden variables we obtain

$$\begin{aligned} p(\bar{\mathbf{X}}) &= \sum_{c_1=1}^{C} \cdots \sum_{c_T=1}^{C} \int_{\mathbf{y}} p(\bar{\mathbf{X}}, \mathbf{y}, Z_1 = c_1, \ldots, Z_T = c_T)\, \mathrm{d}\mathbf{y} \\ &= \sum_{c_1} \cdots \sum_{c_T} \int_{\mathbf{y}} \prod_t p(\mathbf{x}_t|\mathbf{y}, Z_t = c_t) p(Z_t = c_t) p(\mathbf{y})\, \mathrm{d}\mathbf{y} \\ &= \sum_{c_1} \cdots \sum_{c_T} \pi_{c_1} \cdots \pi_{c_T} \mathcal{N}(\bar{\mathbf{X}}; \bar{\boldsymbol{\mu}}_{\bar{\mathbf{c}}}, \bar{\boldsymbol{\Sigma}}_{\bar{\mathbf{c}}}), \end{aligned} \tag{9}$$

where we need to perform a very high number of operations due to all the nested sums which make a total of $C^T$ terms; the vector of indices $\bar{\mathbf{c}} = (c_1, \ldots, c_T)$ contains a particular combination of the indices of the sums; the variable $\bar{\boldsymbol{\mu}}_{\bar{\mathbf{c}}} = (\boldsymbol{\mu}_{c_1}^\intercal, \ldots, \boldsymbol{\mu}_{c_T}^\intercal)^\intercal$ is the concatenated mean super vector corresponding to a particular combination of the indices in the sums; and the covariance matrix has the following structure

$$\bar{\boldsymbol{\Sigma}}_{\bar{\mathbf{c}}} = \bar{\mathbf{W}}_{\bar{\mathbf{c}}}\bar{\mathbf{W}}_{\bar{\mathbf{c}}}^\intercal + \bar{\boldsymbol{\Psi}}_{\bar{\mathbf{c}}}, \tag{10}$$

where $\bar{\mathbf{W}}_{\bar{\mathbf{c}}} = (\mathbf{W}_{c_1}^\intercal, \ldots, \mathbf{W}_{c_T}^\intercal)^\intercal$ is the vertical concatenation of the factor loading matrices also corresponding to $\bar{\mathbf{c}}$, and the internal covariance $\bar{\boldsymbol{\Psi}}_{\bar{\mathbf{c}}}$ is a large diagonal matrix result of the diagonal block composition of the internal covariances for the particular index selection $\bar{\mathbf{c}}$.

To make the notation more compact, we redefine the previous expression as a single sum over all the possible values of $\bar{\mathbf{c}}$ corresponding to the whole sequence as follows

$$p(\bar{\mathbf{X}}) = \sum_{\forall \bar{\mathbf{c}} \in \mathcal{Z}} \bar{\pi}_{\bar{\mathbf{c}}} \mathcal{N}(\bar{\mathbf{X}}; \bar{\boldsymbol{\mu}}_{\bar{\mathbf{c}}}, \bar{\boldsymbol{\Sigma}}_{\bar{\mathbf{c}}}), \tag{11}$$

where $\bar{\pi}_{\bar{\mathbf{c}}} = \prod_t \pi_{c_t}$ is the product of the component weights for a combination of component indices $\bar{\mathbf{c}}$ and $\mathcal{Z} = \{1, \ldots, C\}^T$ is the set of all $C^T$ combinations of $T$ indices with $C$ possible values. We can see that it is equivalent to a mixture of Gaussians with constrained covariance matrices that models the complete

sequence $\bar{\mathbf{X}}$ of dimension $TD$; we will refer to them as super Gaussians. The number of terms in (11) makes exact calculation intractable, but we present here an approximate method to obtain it.

### 3.2. Parameter estimation

First, we present the exact solutions for the E-step, and then, we will focus on the approximations. The expected values of the hidden variables at any iteration of the algorithm are calculated as follows

$$E[\mathbf{y}|\bar{\mathbf{X}}] = \sum_{\forall \bar{\mathbf{c}} \in \mathcal{Z}} p(\bar{\mathbf{c}}|\bar{\mathbf{X}}) E[\mathbf{y}|\bar{\mathbf{X}}, \bar{\mathbf{c}}] \quad (12)$$

$$E[\mathbf{y}\mathbf{y}^\intercal|\bar{\mathbf{X}}] = \sum_{\forall \bar{\mathbf{c}} \in \mathcal{Z}} p(\bar{\mathbf{c}}|\bar{\mathbf{X}}) E[\mathbf{y}\mathbf{y}^\intercal|\bar{\mathbf{X}}, \bar{\mathbf{c}}], \quad (13)$$

where $p(\bar{\mathbf{c}}|\bar{\mathbf{X}})$ is the posterior probability of $\bar{\mathbf{c}}$ given the sequence $\bar{\mathbf{X}}$; and expressions (12) and (13) are first and second order moments of the posterior distribution: $\mathbf{y}|\bar{\mathbf{X}}, \bar{\mathbf{c}}$, the distribution of the latent factors given the observed sequence and component indices $\bar{\mathbf{c}}$. They can be calculated using conditional Gaussian distribution properties [19] as follows

$$E[\mathbf{y}|\bar{\mathbf{X}}, \bar{\mathbf{c}}] = \bar{\mathbf{W}}_{\bar{\mathbf{c}}}^\intercal \bar{\mathbf{\Sigma}}_{\bar{\mathbf{c}}}^{-1} (\bar{\mathbf{X}} - \bar{\boldsymbol{\mu}}_{\bar{\mathbf{c}}}) \quad (14)$$

$$= \mathbf{\Lambda}_{\bar{\mathbf{c}}}^{-1} \bar{\mathbf{W}}_{\bar{\mathbf{c}}}^\intercal \bar{\mathbf{\Psi}}_{\bar{\mathbf{c}}}^{-1} (\bar{\mathbf{X}} - \bar{\boldsymbol{\mu}}_{\bar{\mathbf{c}}}) \quad (15)$$

$$E[\mathbf{y}\mathbf{y}^\intercal|\bar{\mathbf{X}}, \bar{\mathbf{c}}] = \langle \mathbf{y} \rangle \langle \mathbf{y} \rangle^\intercal + \mathbf{I}_R - \bar{\mathbf{W}}_{\bar{\mathbf{c}}}^\intercal \bar{\mathbf{\Sigma}}_{\bar{\mathbf{c}}}^{-1} \bar{\mathbf{W}}_{\bar{\mathbf{c}}} \quad (16)$$

$$= \langle \mathbf{y} \rangle \langle \mathbf{y} \rangle^\intercal + \mathbf{\Lambda}_{\bar{\mathbf{c}}}^{-1}, \quad (17)$$

where $\mathbf{\Lambda}_{\bar{\mathbf{c}}}^{-1} = (\mathbf{I}_R + \bar{\mathbf{W}}_{\bar{\mathbf{c}}}^\intercal \bar{\mathbf{\Psi}}_{\bar{\mathbf{c}}}^{-1} \bar{\mathbf{W}}_{\bar{\mathbf{c}}})^{-1}$; $\langle \mathbf{y} \rangle$ is the expected value (14) in compact notation; the expressions (14) and (16) are hard to calculate due to the high dimension of the super vectors and we have used the matrix inversion lemma to obtain the last expressions.

The expected value for the responsibility latent variables can be found in a similar manner as in the sample independent case, but for sequence super Gaussians

$$p(\bar{\mathbf{c}}|\bar{\mathbf{X}}) = \frac{\bar{\pi}_{\bar{\mathbf{c}}} \mathcal{N}(\bar{\mathbf{X}}; \bar{\boldsymbol{\mu}}_{\bar{\mathbf{c}}}, \bar{\mathbf{\Sigma}}_{\bar{\mathbf{c}}})}{\sum_{\forall \bar{\mathbf{i}} \in \mathcal{Z}} \bar{\pi}_{\bar{\mathbf{i}}} \mathcal{N}(\bar{\mathbf{X}}; \bar{\boldsymbol{\mu}}_{\bar{\mathbf{i}}}, \bar{\mathbf{\Sigma}}_{\bar{\mathbf{i}}})}, \quad (18)$$

where we have to note again the intractability of the sum and that $p(\bar{\mathbf{c}}|\bar{\mathbf{X}})$ is not a product of the posteriors of individual samples.

The M step expressions are identical to the sample independent model [1], since the auxiliary function has the same form given the expected values of the hidden variables. As initialization for the factor loading matrices we perform a singular value decomposition (SVD) using the means of all the components obtained for each utterance [18]. After the new parameters are calculated, we also perform a re-normalization of the parameters of the latent variables —minimum divergence (MD) step [8, 5]— using the observed mean and covariance of the latent variable: $\boldsymbol{\mu}_c^{norm} = \boldsymbol{\mu}_c + \mathbf{W}_c \boldsymbol{\mu}_{\mathbf{y}}$ for the mean vectors; and $\mathbf{W}_c^{norm} = \mathbf{W}_c \mathbf{L}^\intercal$, with $\mathbf{\Sigma}_{\mathbf{y}} = \mathbf{L}\mathbf{L}^\intercal$ the Cholesky decomposition, for the factor loading matrices.

### 3.3. Approximation by sampling

We have shown that exact posterior and likelihood calculations involve sums of $C^T$ terms, which are intractable even for very small number of model components or short sequences. We propose to reduce the number of terms by selecting the most influential of them. The evaluation of the likelihood of a sequence is approximated as

$$p(\bar{\mathbf{X}}) = \sum_{\forall \bar{\mathbf{c}} \in \mathcal{Z}} \bar{\pi}_{\bar{\mathbf{c}}} \mathcal{N}(\bar{\mathbf{X}}; \bar{\boldsymbol{\mu}}_{\bar{\mathbf{c}}}, \bar{\mathbf{\Sigma}}_{\bar{\mathbf{c}}}) \simeq \sum_{\bar{\mathbf{c}}_k \in \mathcal{Z}'_K} \bar{\pi}_{\bar{\mathbf{c}}_k} \mathcal{N}(\bar{\mathbf{X}}; \bar{\boldsymbol{\mu}}_{\bar{\mathbf{c}}_k}, \bar{\mathbf{\Sigma}}_{\bar{\mathbf{c}}_k}), \quad (19)$$

where $\mathcal{Z}'_K = \{\bar{\mathbf{c}}_k | k = 1, \ldots, K, \bar{\mathbf{c}}_k \in \mathcal{Z}\}$ is a subset of $K$ values of $\mathcal{Z}$ and the sum is only evaluated for those values. Note that the cause of the intractability is the number of terms in the sums, since, once the selection is made, the calculations $\mathcal{N}(\bar{\mathbf{X}}; \bar{\boldsymbol{\mu}}_{\bar{\mathbf{c}}_k} \bar{\mathbf{\Sigma}}_{\bar{\mathbf{c}}_k})$ can be obtained very efficiently with the matrix inversion lemma thanks to the matrix covariance structure.

The remaining problem is the selection of the $K$ sequences $\bar{\mathbf{c}}_k$ that provide the highest likelihood. In this work, we only deal with the case $K = 1$ for simplicity, which can be found similar to the idea of training or evaluating a mixture of Gaussians considering only the component with greater responsibility for each sample, but our case is significantly harder since we cannot evaluate all the modes in (19). We rearrange expression (9) as follows

$$p(\bar{\mathbf{X}}) = \int_{\mathbf{y}} p(\mathbf{y}) \sum_{\bar{\mathbf{c}}_k \in \mathcal{Z}} \prod_{t=1}^{T} \pi_{c_t} p(\mathbf{x}_t|\mathbf{y}, Z_t = c_t) \, \mathrm{d}\mathbf{y}$$
$$= \int_{\mathbf{y}} p(\mathbf{y}) \prod_{t=1}^{T} \left( \sum_{c_t=1}^{C} \pi_{c_t} p(\mathbf{x}_t|\mathbf{y}, Z_t = c_t) \right) \mathrm{d}\mathbf{y}, \quad (20)$$

where $p(\mathbf{x}_t|\mathbf{y}, Z_t = c_t)$ follows the distribution (7). We can see in the factorization (20) that the likelihood of the sequence given the latent variable is equal to the product of MFA like models, which can also be deduced from the Bayesian network.

The previous expression is more convenient to discover the most important component assignments, since the super Gaussian covariance matrix has a factor analysis structure (10); therefore when used in applications where data exhibit dominant principal vectors —like speaker recognition—, responsibilities are more sensitive to changes in those dimensions of the latent variables. Then, we propose to sample only those directions of the hidden variable, which can be done efficiently, to find the best responsibilities for a sequence

$$p(\bar{\mathbf{X}}) \simeq \sum_{\tilde{y}_1, \ldots, \tilde{y}_{\tilde{R}}} p(\tilde{\mathbf{y}}) \prod_{t=1}^{T} \left( \sum_{c_t=1}^{C} \pi_{c_t} p(\mathbf{x}_t|\tilde{\mathbf{y}}, Z_t = c_t) \right), \quad (21)$$

where $\tilde{\mathbf{y}} \in \{y_{min}, \ldots, y_{max}\}^{\tilde{R}}$ is a sampled version of the latent space with dimension $\tilde{R}$ very small and tractable (in this work $\tilde{R} = 2$). Then, the variable $\tilde{\mathbf{y}}$ is evaluated in a finite number of points in the range $[y_{min}, y_{max}]$ in this work the number of points is set to 15 and the range is $[-2.5, 2.5]$. The approximate solution is obtained in two steps, first we choose the highest term in the sum over the subspace of the sampled latent variable $\tilde{\mathbf{y}}$. Once the best value of $\tilde{\mathbf{y}}$ is found we obtain the best components $c_t$ and compose the best sequence $\bar{\mathbf{c}}_k$, since it can now be solved independently for each component. Empirically, we have observed the bigger terms found in the sampling are usually several orders of magnitude above the following ones.

### 3.4. Additional levels of hidden variable structure

In the same way as JFA is done [6, 7], these models can also be used to deal with several layers of tied latent variables. For

example in this work we used two levels of hidden variables to model frames coming from the same speaker (speaker variability) and frames coming from the same audio clip (channel variability). For $N$ sequences of the same speaker, this can be formalized by concatenating all the hidden variables as $\mathbf{y} = (\mathbf{y}^{(2)\intercal}, \mathbf{y}_1^{(1)\intercal}, \ldots, \mathbf{y}_N^{(1)\intercal})^\intercal$, where $\mathbf{y}^{(2)}$ is the speaker factor and $\mathbf{y}_n^{(1)}$ is the channel factor of utterance $n$, with sizes $R^{(2)}$ and $R^{(1)}$ respectively. For examples, in this work the number of enrollment utterances per speaker is $N = 3$. Then, to compute the E step posteriors and data likelihoods we can construct the new factor loading super matrix, $\bar{\mathbf{W}}_{\bar{c}}$, which an illustrative example of $N = 2$ utterances and length $T$ would have the following structure

$$\bar{\mathbf{W}}_{\bar{c}} = \begin{pmatrix} \mathbf{W}_{c_{11}}^{(2)} & \mathbf{W}_{c_{11}}^{(1)} & \\ \vdots & \vdots & \\ \mathbf{W}_{c_{1T}}^{(2)} & \mathbf{W}_{c_{1T}}^{(1)} & \\ \mathbf{W}_{c_{21}}^{(2)} & & \mathbf{W}_{c_{21}}^{(1)} \\ \vdots & & \vdots \\ \mathbf{W}_{c_{2T}}^{(2)} & & \mathbf{W}_{c_{2T}}^{(1)} \end{pmatrix}. \tag{22}$$

# 4. Experiments

The experiments have been conducted on the RSR2015 text dependent speaker recognition database [20, 21], composed of 9 utterances of 30 different phrases recorded in different sessions by 299 speakers. The speakers are distributed in three groups: background, development and evaluation. We have used only background data to train our universal background models (UBM) which can be phrase independent or phrase dependent, as in [12]. The evaluation part is used for enrollment and testing speaker models. Speakers are enrolled using 3 utterances of each phrase and for testing we have selected the trials using the same phrase as the model.

The UBMs have been trained with the background partition using MFA models and Gaussian mixtures for the baseline and the proposed TMFA models. Two types of experiments are defined depending on the UBM. The first set of experiments are performed using gender dependent and phrase independent UBMs. In the second set we used phrase and gender dependent UBMs; therefore, they become more specific phonetically but the amount of data available to train them is smaller. The features are 19 Mel filterbank cepstrum coefficients (MFCC) plus the log energy and two derivatives are taken, providing a total of 60 dimensions obtained from 25ms windows with advance of 10ms. Then, an energy based voice activity detector is used and data are normalized using short term gaussianization.

We have obtained speaker recognition scores by using log-likelihood ratios (LLR), defined as the log of the $p(\bar{\mathbf{X}}|\mathbf{\Theta}^{SPK})$ to $p(\bar{\mathbf{X}}|\mathbf{\Theta}^{UBM})$ ratio, where the likelihoods are calculated using (5) for MFAs or (19) for TMFAs; $\mathbf{\Theta}^{UBM}$ are the UBM parameters; and $\mathbf{\Theta}^{SPK}$ are the speaker dependent parameters. For enrollment we used a maximum a posteriori (MAP)[22] over the mean parameters of the model with $\tau = 1$ with one iteration, which is equivalent to $dz$ term in JFA. For TMFAs, we applied the MD re-normalization using observed statistics of the latent variables to the speaker mean vectors and factor loading matrices (as expressed in Section 3); therefore, the means are updated using the speaker mean latent factor; and factor loading matrices are also normalized so that the Gaussian equivalent covariances are now dominated by the speaker intrinsic variability $\mathbf{\Psi}_c$ (unaltered from the UBM).

Table 1: Experimental results on RSR2015 [20] eval set with same phrase trials, where EER% and NIST 2008 and 2010 min costs (det08,det10) are shown.

| | | Male | | | Female | | |
|---|---|---|---|---|---|---|---|
| System | $C$ | EER% | det08 | det10 | EER% | det08 | det10 |
| MGauss | 512 | 0.93 | 0.049 | 0.192 | 1.43 | 0.061 | 0.188 |
| MFA (25) | 64 | 1.99 | 0.109 | 0.440 | 2.48 | 0.124 | 0.436 |
| | 128 | 1.54 | 0.088 | 0.383 | 1.90 | 0.090 | 0.313 |
| | 256 | 1.42 | 0.075 | 0.310 | 1.53 | 0.071 | 0.244 |
| | 512 | 1.23 | 0.065 | 0.270 | **1.23** | 0.059 | 0.205 |
| TMFA$^\star$ (25,5) | 64 | 1.25 | 0.068 | 0.291 | 2.15 | 0.097 | 0.280 |
| TMFA (25,5) | 64 | 0.84 | 0.042 | **0.154** | 1.32 | 0.056 | 0.174 |
| | 128 | **0.81** | **0.041** | 0.161 | 1.24 | **0.046** | **0.131** |
| MFA$^{pd}$ (25) | 64 | 1.38 | 0.075 | 0.337 | 0.96 | 0.047 | 0.198 |
| TMFA$^{\star pd}$ (25,5) | 64 | **0.48** | **0.028** | 0.138 | 0.76 | 0.029 | 0.099 |
| TMFA$^{pd}$ (25,5) | 64 | 0.61 | 0.030 | **0.123** | **0.55** | **0.023** | **0.088** |

In Table 1 we can see results of the equal error rate (EER) and NIST 2008 and 2010 minimum detection cost functions for female and male subsets comparing mixture of Gaussians with diagonal covariance matrices, MFA and TMFA models. All results shown in the table are obtained from LLRs without score normalization. We have found that TMFA models can provide interesting results for a number of speaker factors, channel factors and mixture components smaller than usual factor analysis systems. In these experiments the latent factor dimensions are ($R^{(2)} = 25$, $R^{(1)} = 5$). According to the costs in Table 1, it can be seen that TMFA presents a very good behavior in the low false alarm region. We also present results of TMFA models in which the evaluation is done frame independently —called TMFA$^\star$— showing better performance than standard MFA.

In the lower part of Table 1 we present results using the phrase dependent UBMs, which reduce phonetic variability. In this case the difference between TMFA and TMFA$^\star$ is smaller, which can be due to the reduced phonetic variability; nevertheless, TMFA training again provides better results than standard MFA.

# 5. Conclusions

In this paper we present a new method to model underlying information shared in sequences based on the assumption that latent variables that explain it are tied for a number of frames. The model explains sequence and inter-sequence common information as hidden factors, for which we present both exact —but intractable— and approximate methods based on sampling to estimate model parameters, evaluate posterior probabilities and evaluate likelihoods. The generalization ability of these factorizations has been previously addressed in many works [4, 5]; however, the proposed model differs in the inference of hidden variables, parameter estimation and likelihood evaluation. We have tested the models in the text dependent speaker recognition database RSR2015. Results confirm TMFA estimation by sampling as an interesting line since they have been obtained by using a simple approximation to the exact solution and small size models.

# 6. Acknowledgements

This work was funded by the Spanish Government and the European Union (FEDER) under project TIN2011-28169-C05-02.

# 7. References


[1] Z. Ghahramani and G. Hinton, "The EM algorithm for mixtures of factor analyzers," Dept. of Comp. Sci., Univ. of Toronto, Toronto, Tech. Rep. 1, 1996.

[2] T. Hasan and J. H. L. Hansen, "Acoustic Factor Analysis for Robust Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 842–853, 2013.

[3] ——, "Maximum Likelihood Acoustic Factor Analysis Models for Robust Speaker Verification in Noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 381–391, Feb. 2014.

[4] S.-C. Yin, R. Rose, and P. Kenny, "A Joint Factor Analysis Approach to Progressive Model Adaptation in Text-Independent Speaker Verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1999–2010, Sep. 2007.

[5] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, Jul. 2008.

[6] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[7] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, May 2005.

[8] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability : Theory and Algorithms," CRIM, Montreal, CRIM-06/08-13, Tech. Rep., 2005.

[9] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification," in *in Proceedings of Interspeech*, Brighton, UK, Sep 2009.

[10] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis For Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 14, pp. 788–798, May 2010.

[11] J. Villalba and N. Brümmer, "Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance," in *Interspeech 2011*, Florence, 2011, pp. 28–31.

[12] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-dependent speaker recogntion using plda with uncertainty propagation," in *in Proceedings of Interspeech*, Lyon, France, August 2013.

[13] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brasil, 2007.

[14] D. Martínez, O. Plchot, L. Burget, G. Ondrej, and P. Matejka, "Language Recognition in iVectors Space," in *in Proceedings of Interspeech*, Florence, Italy, 2011.

[15] D. Martínez, L. Burget, L. Ferrer, and N. Scheffer, "iVector-Based Prosodic System for Language Identification," in *ICASSP*, Kyoto, Japan, 2012.

[16] D. Castan, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Segmentation-by-Classification System Based on Factor Analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[17] D. Castan, A. Ortega, A. Miguel, and E. Lleida, "Broadcast News Segmentation with Factor Analysis System," in *SLAM Workshop*, 2013, pp. 1–6.

[18] C. Vaquero, A. Ortega, A. Miguel, and E. Lleida, "Quality Assessment for Speaker Diarization and Its Application in Speaker Characterization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 816–827, Apr. 2013.

[19] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.

[20] A. Larcher, K. A. Lee, B. Ma, and H. Li, "The RSR2015 database for text-dependent speaker verification using multiple pass-phrases," in *in Proceedings of Interspeech*, Portland (Oregon), USA, Sept 2012.

[21] ——, "Text-dependent Speaker Verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[22] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2(2), pp. 291–298, 1994.