Mixtures of Bayesian Joint Factor Analyzers for Noise Robust Automatic Speech Recognition

Xiaodong Cui, Vaibhava Goel and Brian Kingsbury

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

{cuix,vgoel,bedk}@us.ibm.com

Abstract

This paper investigates a noise robust approach to automatic speech recognition based on a mixture of Bayesian joint factor analyzers. In this approach, noisy features are modeled by two joint groups of factors accounting for speaker and noise variabilities which are estimated by clean and noisy speech respectively. The factors form an overcomplete dictionary with a redundant representation. Automatic relevance determination (ARD) is carried out by the relevance vector machine (RVM) where sparsity-promoting priors are applied on two factor loading matrices. Experiments on large vocabulary continuous speech recognition (LVCSR) tasks show good improvements by this approach.

Index Terms: Bayesian joint factor analysis, automatic relevance determination, relevance vector machine, noise robustness, LVCSR

1. Introduction

Joint factor analysis (JFA), as an extended form of the classical factor analysis (FA)[12][13], aims to interpret the observed signals by multiple distinct groups of underlying factors. In recent years, it finds its success in a variety of speaker recognition and verification tasks [1][2][3][4] where JFA is used to model speaker and channel variabilities and yields the state-of-the-art performance. In those applications, JFA is conducted in a very high dimensional space of supervectors. The total number of underlying factors is smaller than the dimension of the supervectors and is usually determined by heuristics.

In this paper, we investigates a feature compensation scheme based on joint factor analysis for noise robust automatic speech recognition (ASR). Specifically, we model the noisy features by a mixture of Bayesian joint factor analyzers where each joint factor analyzer has two groups of factors to explicitly describe speaker and noise variabilities. The whole set of factors creates an overcomplete dictionary for a redundant representation of the noisy features. Automatic relevance determination (ARD) [7] is carried out to automatically learn the effective underlying factors given the training data by the relevance vector machine (RVM) [5] where sparsity-promoting priors are applied to the two factor loading matrices.

In this Bayesian framework, the hyper-parameters of the prior distributions are estimated under the type-II maximum likelihood (ML) criterion [7] while ML estimates of the other parameters are obtained by the Expectation-Maximization (EM) algorithm [6]. The factors for speaker variability are learned from clean speech while the factors for noise variability are learned from noisy speech given the estimated speaker factors. In particular, the two groups of factors are initialized by probabilistic principal component analysis (PPCA) [8], which corresponds to eigen-voice and eigen-noise, respectively.

In the noise compensation stage, both speaker and noise variabilities are estimated based on the whole factor dictionary and the noise component is removed as the undesired variability from the input noisy speech.

The remainder of the paper is organized as follows. Section 2 presents the mathematical formulation of the proposed mixture of Bayesian joint factor analyzers. Section 3 gives the details of the stage-wise parameter estimation of the two groups of factors using clean and noisy speech. Section 4 describes how the noise compensation is performed to recover the clean speech. Experimental results on large vocabulary continuous speech recognition (LVCSR) are reported in Section 5.

2. Mixture of Bayesian Joint Factor Analyzers

Let $\boldsymbol{y}$ be the observation vector. Define a mixture of factor analyzers

$$\boldsymbol{y} = \sum_{k=1}^{K} c_k \varphi_k(\boldsymbol{y}) \tag{1}$$

where c_k are the mixture weights and $\varphi_k(\boldsymbol{y})$ is a factor analyzer in component k which has the following form

$$\varphi_k(\boldsymbol{y}) = \boldsymbol{\mu}_k + \mathbf{A}_k \boldsymbol{\xi}_k + \mathbf{B}_k \boldsymbol{\eta}_k + \boldsymbol{\epsilon}_k. \tag{2}$$

In Eq.2, $\mathbf{A}_k$ and $\mathbf{B}_k$, which are known as factor loading matrices in the FA literature, are referred to as factor dictionaries in this paper. They are used to describe distinct speech variabilities. Specifically, in the context of noise robustness discussed here where $\boldsymbol{y}$ are noisy speech features, $\mathbf{A}_k$ and $\mathbf{B}_k$ are for speaker and noise variabilities, respectively. $\boldsymbol{\xi}_k$ and $\boldsymbol{\eta}_k$ are latent variables and $\boldsymbol{\epsilon}_k$ is a random variable for the residual. They all obey Gaussian distributions:

$$p(\boldsymbol{\xi}_k) \sim \mathcal{N}(\boldsymbol{\xi}_k; \mathbf{0}, \mathbf{I}) \tag{3}$$

$$p(\boldsymbol{\eta}_k) \sim \mathcal{N}(\boldsymbol{\eta}_k; \mathbf{0}, \mathbf{I}) \tag{4}$$

$$p(\boldsymbol{\epsilon}_k) \sim \mathcal{N}(\boldsymbol{\epsilon}_k; \mathbf{0}, \boldsymbol{\Psi}_k) \tag{5}$$

where $\boldsymbol{\Psi}_k$ is a diagonal matrix.

Define

$$\mathbf{W}_k = [\mathbf{A}_k \ \mathbf{B}_k] \quad \text{and} \quad \boldsymbol{\zeta}_k = [\boldsymbol{\xi}_k^{\mathsf{T}} \ \boldsymbol{\eta}_k^{\mathsf{T}}]^{\mathsf{T}} \tag{6}$$

Eq.2 can be rewritten as

$$\varphi_k(\boldsymbol{y}) = \boldsymbol{\mu}_k + \mathbf{W}_k \boldsymbol{\zeta}_k + \boldsymbol{\epsilon}_k \tag{7}$$

and $p(\boldsymbol{\zeta}_k) \sim \mathcal{N}(\boldsymbol{\zeta}_k; \mathbf{0}, \mathbf{I})$.

Integrating out the latent variable ζ_k and the residual variable $\boldsymbol{\epsilon}_k$, the distribution of $\boldsymbol{y}$ in mixture component k is a Gaussian

$$p_k(\boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{8}$$

where

$$\boldsymbol{\Sigma}_k = \mathbf{W}_k\mathbf{W}_k^{\mathsf{T}} + \boldsymbol{\Psi}_k = \mathbf{A}_k\mathbf{A}_k^{\mathsf{T}} + \mathbf{B}_k\mathbf{B}_k^{\mathsf{T}} + \boldsymbol{\Psi}_k \tag{9}$$

Accordingly, the mixture of factor analyzers in Eq.1 is equivalent to a Gaussian mixture model (GMM)

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} c_k\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{10}$$

The overall factor dictionary $\mathbf{W}_k$ in Eq.7 is chosen to be overcomplete and the RVM is used to learn the sparse underlying factors that can effectively interpret the training data. Given the way $\mathbf{A}_k$ and $\mathbf{B}_k$ are separately estimated using clean and noisy speech, the following prior distributions are applied on $\mathbf{A}_k$ and $\mathbf{B}_k$

$$p(\mathbf{A}_k|\alpha_k^{(\mathbf{A})}) = \prod_{i=1}^{d-1}\left(\frac{\alpha_{k,i}^{(\mathbf{A})}}{2\pi}\right)^{\frac{d}{2}} \exp\left\{-\frac{1}{2}\alpha_{k,i}^{(\mathbf{A})}\mathbf{a}_{k,i}^{\mathsf{T}}\mathbf{a}_{k,i}\right\} \tag{11}$$

$$p(\mathbf{B}_k|\alpha_k^{(\mathbf{B})}) = \prod_{i=1}^{d-1}\left(\frac{\alpha_{k,i}^{(\mathbf{B})}}{2\pi}\right)^{\frac{d}{2}} \exp\left\{-\frac{1}{2}\alpha_{k,i}^{(\mathbf{B})}\mathbf{b}_{k,i}^{\mathsf{T}}\mathbf{b}_{k,i}\right\} \tag{12}$$

where $\mathbf{a}_{k,i}$ and $\mathbf{b}_{k,i}$ are the columns of factor dictionaries $\mathbf{A}_k$ and $\mathbf{B}_k$ and d is the dimensionality of the input features. The numbers of columns of both dictionaries are equal to $d-1$ for a reason that will become clear shortly. Based on the RVM theory [5], if an atom (i.e. a column) $\mathbf{w}_i$ in the dictionary is poorly aligned with the training data, its hyper-parameter α_i will be driven to infinity and this atom will be effectively switched off.

3. Parameter Estimation

Since we want to explicitly use $\mathbf{A}_k$ to describe speaker variability and $\mathbf{B}_k$ to describe noise variability, the parameter estimation is conducted in a stage-wise manner. The speaker factor dictionary $\mathbf{A}_k$ is first estimated using the clean speech by nulling the noise factor dictionary $\mathbf{B}_k$. Given the estimated speaker factor dictionary $\mathbf{A}_k$, the noise factor dictionary $\mathbf{B}_k$ is then estimated using the noisy speech.

3.1. Estimation of $\mathbf{A}_k$

Let $\boldsymbol{x}$ be the observed clean features. To estimate the speaker dictionary $\mathbf{A}_k$ in component k, the noise dictionary $\mathbf{B}_k$ is nulled and Eq.2 becomes

$$\varphi_k(\boldsymbol{x}) = \boldsymbol{\mu}_{k,a} + \mathbf{A}_k\boldsymbol{\xi}_k + \boldsymbol{\epsilon}_k. \tag{13}$$

The ML estimation is initialized by assuming an isotropic residual covariance

$$\boldsymbol{\Psi}_k = \sigma_k^2\mathbf{I} \tag{14}$$

which corresponds to a PPCA setting [8]. It has a closed-form solution

$$\mu_{k,a} = \frac{\sum_{t=1}^{T}\gamma_k(t)\boldsymbol{x}_t}{\sum_{t=1}^{T}\gamma_k(t)}, \quad \sigma_k^2 = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_{k,j} \tag{15}$$

$$\mathbf{A}_k = \mathbf{U}_{k,q}(\boldsymbol{\Lambda}_{k,q} - \sigma_k^2\mathbf{I})^{1/2} \tag{16}$$

where $\{\lambda_{k,1}, \cdots, \lambda_{k,d}\}$ are the eigenvalues, in the decreasing order, of the sample covariance of component k

$$\mathbf{S}_k = \frac{\sum_{t=1}^{T}\gamma_k(t)(\boldsymbol{x}_t - \boldsymbol{\mu}_{k,a})(\boldsymbol{x}_t - \boldsymbol{\mu}_{k,a})^{\mathsf{T}}}{\sum_{t=1}^{T}\gamma_k(t)} \tag{17}$$

with T being the total number of samples and $\gamma_k(t)$ the posterior probability of component k at time t. Matrix $\mathbf{U}_{k,q}$ has q eigenvectors as its columns corresponding to the first q most significant eigenvalues $\{\lambda_{k,1}, \cdots, \lambda_{k,q}\}$ and $\boldsymbol{\Lambda}_{k,q} = \text{diag}\{\lambda_{k,1}, \cdots, \lambda_{k,q}\}$. The estimate of σ_k^2 is simply the sum of the rest of the $d-q$ eigenvalues divided by $d-q$. The number of factors q is set to $d-1$ which is the maximum value allowed by PPCA. Given the way $\mathbf{A}_k$ are computed in Eq.16, it is equivalent to eigen-speaker or eigen-voice modeling in a probabilistic linear-Gaussian form.

After the eigen-speaker initialization, the speaker factor dictionary $\mathbf{A}_k$ are then estimated by removing the isotropic constraints in Eq.14. There is no closed-form solution in this case and the parameters can be updated iteratively. First of all, it can be shown that the hyper-parameter of the prior distribution can be updated by the type-II ML criterion [7][9]

$$\widetilde{\alpha}_{k,i}^{(\mathbf{A})} \approx \frac{d}{||\mathbf{a}_{k,i}||^2} \tag{18}$$

Other parameters can be estimated by the EM algorithm [7].

In the E-step, compute

$$\mathbf{E}[\boldsymbol{\xi}_{k,t}] = \mathbf{G}_k\mathbf{A}_k^{\mathsf{T}}\boldsymbol{\Psi}_k^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_{k,a}) \tag{19}$$

$$\mathbf{E}[\boldsymbol{\xi}_{k,t}\boldsymbol{\xi}_{k,t}^{\mathsf{T}}] = \mathbf{G}_k + \mathbf{E}[\boldsymbol{\xi}_{k,t}]\mathbf{E}[\boldsymbol{\xi}_{k,t}]^{\mathsf{T}} \tag{20}$$

where $\mathbf{G}_k = (\mathbf{I} + \mathbf{A}_k^{\mathsf{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{A}_k)^{-1}$.

In the M-step, first define

$$\mathbf{L}_k \triangleq \frac{\sum_{t=1}^{T}\gamma_k(t)\mathbf{E}[\boldsymbol{\xi}_{k,t}\boldsymbol{\xi}_{k,t}{}^{\mathsf{T}}]}{\sum_{t=1}^{T}\gamma_k(t)} \tag{21}$$

$$= \mathbf{G}_k + \mathbf{G}_k\mathbf{A}_k^{\mathsf{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{S}_k\boldsymbol{\Psi}_k^{-1}\mathbf{A}_k\mathbf{G}_k \tag{22}$$

$$\mathbf{R}_k \triangleq \frac{\sum_{t=1}^{T}\gamma_k(t)(\boldsymbol{x}_t - \boldsymbol{\mu}_{k,a})\mathbf{E}[\boldsymbol{\xi}_{k,t}]^{\mathsf{T}}}{\sum_{t=1}^{T}\gamma_k(t)} \tag{23}$$

$$= \mathbf{S}_k\boldsymbol{\Psi}_k^{-1}\mathbf{A}_k\mathbf{G}_k \tag{24}$$

$$\boldsymbol{\Psi}_k = \textbf{diag}\{\sigma_{k,1}^2, \cdots, \sigma_{k,d}^2\} \tag{25}$$

$$\boldsymbol{\Delta}_k = \textbf{diag}\{\alpha_{k,1}^{(\mathbf{A})}, \cdots, \alpha_{k,d-1}^{(\mathbf{A})}\} \tag{26}$$

It can be shown that the parameters are updated as follows [9]

$$\widetilde{c}_k = \frac{1}{T}\sum_{t=1}^{T}\gamma_k(t), \quad \widetilde{\boldsymbol{\mu}}_{k,a} = \frac{\sum_{t=1}^{T}\gamma_k(t)\boldsymbol{x}_t}{\sum_{t=1}^{T}\gamma_k(t)} \tag{27}$$

$$\widetilde{\bar{\boldsymbol{a}}}_{k,i}^{(\mathbf{A})} = \bar{\mathbf{r}}_{k,i}\left(\frac{\sigma_{k,i}^2}{\sum_{t=1}^{T}\gamma_k(t)}\boldsymbol{\Delta}_k + \mathbf{L}_k\right)^{-1}, \quad i = 1, \cdots, d \tag{28}$$

$$\widetilde{\boldsymbol{\Psi}}_k = \text{diag}\left\{\mathbf{S}_k - \widetilde{\mathbf{A}}_k\mathbf{G}_k\mathbf{A}_k^{\mathsf{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{S}_k\right\} \tag{29}$$

where $\bar{\boldsymbol{a}}$ and $\bar{\mathbf{r}}$ are rows of $\mathbf{A}$ and $\mathbf{R}$.

3.2. Estimation of $\mathbf{B}_k$

When the speaker factor dictionary $\mathbf{A}_k$ is learned, $\mathbf{B}_k$ is reintroduced back to Eq.2 and estimated for each component using the noisy speech. To initialize, the noise component is first

predicted from the noisy features. Given an observed noisy feature $\boldsymbol{y}$, the corresponding clean speech component is predicted in component k as

$$\hat{\boldsymbol{x}}_k = \boldsymbol{\mu}_{k,a} + \mathbf{A}_k \mathbf{E}[\boldsymbol{\xi}_k|\boldsymbol{y}] \tag{30}$$

$$= \boldsymbol{\mu}_{k,a} + \mathbf{A}_k \mathbf{G}_k \mathbf{A}_k^{\mathsf{T}} \boldsymbol{\Psi}_k^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_{k,a}) \tag{31}$$

and accordingly the noise component is

$$\boldsymbol{n}_k = \boldsymbol{y} - \hat{\boldsymbol{x}}_k \tag{32}$$

Following a similar procedure from Sec. 3.1, an isotropic residual covariance matrix is first assumed as Eq.14 and $\mathbf{B}_k$ are computed in the PPCA framework as

$$\boldsymbol{\mu}_k = \frac{\sum_{t=1}^T \gamma_k(t) \boldsymbol{y}_t}{\sum_{t=1}^T \gamma_k(t)}, \quad \boldsymbol{\mu}_{k,b} = \boldsymbol{\mu}_k - \boldsymbol{\mu}_{k,a} \tag{33}$$

$$\mathbf{B}_k = \mathbf{U}_{k,d-1} (\boldsymbol{\Lambda}_{k,d-1} - \sigma_k^2 \mathbf{I})^{1/2}, \quad \sigma_k^2 = \lambda_{k,d} \tag{34}$$

where the number of the columns of $\mathbf{B}_k$ is also set to $d-1$ and $\mathbf{U}_{k,d-1}$ is composed of $d-1$ eigenvectors of the sample covariance of $\boldsymbol{n}_k$ in component k

$$\mathbf{S}_k = \frac{\sum_{t=1}^T \gamma_k(t) (\boldsymbol{n}_{k,t} - \boldsymbol{\mu}_{k,b})(\boldsymbol{n}_{k,t} - \boldsymbol{\mu}_{k,b})^{\mathsf{T}}}{\sum_{t=1}^T \gamma_k(t)} \tag{35}$$

as its columns corresponding to the $d-1$ most significant eigenvalues $\{\lambda_{k,1}, \cdots, \lambda_{k,d-1}\}$ in decreasing order. Therefore, the noise factor dictionary initialized this way is equivalent to an eigen-noise model setting.

Once the noise factor dictionary has been initialized, the parameters are updated iteratively. Following a similar derivation of Eq.18, one has the type-II ML estimate of the hyperparameter

$$\widetilde{\alpha}_{k,i}^{(\mathbf{B})} \approx \frac{d}{||\mathbf{b}_{k,i}||^2} \tag{36}$$

The remaining parameters are estimated by the EM algorithm.

In the E-step, define the auxiliary function

$$\mathbf{Q}(\lambda, \bar{\lambda}) = \mathbf{E}_{\boldsymbol{Z}} \mathbf{E}_{\boldsymbol{\Omega}} [\log p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Omega}) p(\mathbf{W}|\alpha) | \boldsymbol{Y}, \lambda)] \tag{37}$$

$$= \sum_{t=1}^T \sum_{k=1}^K \gamma_k(t) \mathbf{E}_{\boldsymbol{Z}} [\log p(\boldsymbol{y}_t, \boldsymbol{\zeta}_t, k | \boldsymbol{Y}, \lambda) p(\mathbf{W}|\alpha)] \tag{38}$$

where $\boldsymbol{Y} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T\}$ are noisy observation vectors; $\boldsymbol{Z} = \{\boldsymbol{\zeta}_1, \cdots, \boldsymbol{\zeta}_T\}$ are latent factor dictionary weighting coefficients; $\boldsymbol{\Omega} = \{\omega_1, \cdots, \boldsymbol{\omega}_T\}$ are latent component indices in the mixture. To estimate $\mathbf{B}_k$, express the auxiliary function with respect to $\mathbf{B}_k$

$$\mathbf{Q}_{\mathbf{B}_k}(\lambda, \bar{\lambda}) = \sum_{t=1}^T \sum_{k=1}^K \gamma_k(t) \Big\{ 2\mathbf{Tr}(\boldsymbol{\Psi}_k^{-1} \mathbf{A}_k \mathbf{E}[\boldsymbol{\xi}_{k,t} \boldsymbol{\eta}_{k,t}^{\mathsf{T}}] \mathbf{B}_k^{\mathsf{T}})$$

$$+ \mathbf{Tr}(\boldsymbol{\Psi}_k^{-1} \mathbf{B}_k \mathbf{E}[\boldsymbol{\eta}_{k,t} \boldsymbol{\eta}_{k,t}^{\mathsf{T}}] \mathbf{B}_k^{\mathsf{T}}) - 2(\boldsymbol{y}_t - \boldsymbol{\mu}_k)^{\mathsf{T}} \boldsymbol{\Psi}_k^{-1} (\mathbf{B}_k \mathbf{E}[\boldsymbol{\eta}_{k,t}]) \Big\}$$

$$+ \sum_{i=1}^{d-1} \alpha_{k,i}^{(\mathbf{B})} \boldsymbol{b}_{k,i}^{\mathsf{T}} \boldsymbol{b}_{k,i} + \text{const} \tag{39}$$

Taking the derivative with respect to $\mathbf{B}_k$ and setting it to zero, one has

$$\boldsymbol{\Psi} \mathbf{B}_k \boldsymbol{\Delta}_k + \mathbf{B}_k \sum_{t=1}^T \gamma_k(t) \mathbf{E}[\boldsymbol{\eta}_{k,t} \boldsymbol{\eta}_{k,t}^{\mathsf{T}}] =$$

$$\sum_{t=1}^T \gamma_k(t) (\boldsymbol{y}_t - \boldsymbol{\mu}_k) \mathbf{E}[\boldsymbol{\eta}_{k,t}^{\mathsf{T}}] - \sum_{t=1}^T \gamma_k(t) \mathbf{A}_k \mathbf{E}[\boldsymbol{\xi}_{k,t} \boldsymbol{\eta}_{k,t}^{\mathsf{T}}] \tag{40}$$

Define the following matrices

$$\mathbf{L}_k \triangleq \frac{\sum_{t=1}^T \gamma_k(t) \mathbf{E}[\boldsymbol{\eta}_{k,t} \boldsymbol{\eta}_{k,t}^{\mathsf{T}}]}{\sum_{t=1}^T \gamma_k(t)} \tag{41}$$

$$\mathbf{R}_k \triangleq \frac{\sum_{t=1}^T \gamma_k(t) (\boldsymbol{y}_t - \boldsymbol{\mu}_k) \mathbf{E}[\boldsymbol{\eta}_{k,t}^{\mathsf{T}}] - \sum_{t=1}^T \gamma_k(t) \mathbf{A}_k \mathbf{E}[\boldsymbol{\xi}_{k,t} \boldsymbol{\eta}_{k,t}^{\mathsf{T}}]}{\sum_{t=1}^T \gamma_k(t)} \tag{42}$$

Since $\boldsymbol{\Psi}_k$ is diagonal, Eq.40 can be solved row by row. Let $\bar{\boldsymbol{b}}_{k,i}$ be the ith row of the dictionary $\mathbf{B}_k$ and $\bar{\boldsymbol{r}}_{k,i}$ be the ith row of matrix $\mathbf{R}_k$, then Eq.40 can be expressed as

$$\begin{bmatrix} \bar{\boldsymbol{b}}_{k,1} \left(\frac{\sigma_{k,1}^2}{\sum_{t=1}^T \gamma_k(t)} \boldsymbol{\Delta}_k + \mathbf{L}_k \right) \\ \vdots \\ \bar{\boldsymbol{b}}_{k,d} \left(\frac{\sigma_{k,d}^2}{\sum_{t=1}^T \gamma_k(t)} \boldsymbol{\Delta}_k + \mathbf{L}_k \right) \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{r}}_{k,1} \\ \vdots \\ \bar{\mathbf{r}}_{k,d} \end{bmatrix} \tag{43}$$

It follows that

$$\widetilde{\bar{\boldsymbol{b}}}_{k,i} = \bar{\mathbf{r}}_{k,i} \left(\frac{\sigma_{k,i}^2}{\sum_{t=1}^T \gamma_k(t)} \boldsymbol{\Delta}_k + \mathbf{L}_k \right)^{-1}, \quad i = 1, \cdots, d \tag{44}$$

where

$$\boldsymbol{\Delta}_k = \mathbf{diag}\{\alpha_{k,1}^{(\mathbf{B})}, \cdots, \alpha_{k,d-1}^{(\mathbf{B})}\} \tag{45}$$

Analogously, taking the derivative of the auxiliary function with respect to $\boldsymbol{\Psi}_k$ and setting it to zero, one has

$$\boldsymbol{\Psi}_k = \mathbf{diag} \Big\{ \mathbf{S}_k - \frac{1}{\sum_{t=1}^T \gamma_k(t)} \Big[2\mathbf{W}_k \sum_{t=1}^T \gamma_k(t) \mathbf{E}[\boldsymbol{\zeta}_{k,t}] (\boldsymbol{y}_t - \boldsymbol{\mu}_k)^{\mathsf{T}}$$

$$- \mathbf{W}_k \sum_{t=1}^T \gamma_k(t) \mathbf{E}[\boldsymbol{\zeta}_{k,t} \boldsymbol{\zeta}_{k,t}^{\mathsf{T}}] \mathbf{W}_k^{\mathsf{T}} \Big] \Big\} \tag{46}$$

$$= \mathbf{diag} \Big\{ \mathbf{S}_k - 2\mathbf{W}_k \mathbf{G}_k \mathbf{W}_k^{\mathsf{T}} \boldsymbol{\Psi}_k^{-1} \mathbf{S}_k + \mathbf{W}_k (\mathbf{G}_k +$$

$$\mathbf{G}_k \mathbf{W}_k^{\mathsf{T}} \boldsymbol{\Psi}_k^{-1} \mathbf{S}_k \boldsymbol{\Psi}_k^{-1} \mathbf{W}_k \mathbf{G}_k) \mathbf{W}_k^{\mathsf{T}} \Big\} \tag{47}$$

The updates of mean $\boldsymbol{\mu}_b$ and mixture weight c_k are the same as Eq.33 and Eq.27, respectively.

4. Noise Compensation

After the model parameters are in place, the predictive distribution of $\boldsymbol{y}$ is a GMM distribution under the assumption of a sharp peak around the MAP estimate of $\mathbf{W}_k^{\mathsf{MAP}}$. Given the input noisy features, $\boldsymbol{y}_t$, the noise variability can be estimated in each mixture component k as

$$\boldsymbol{n}_{k,t} = \boldsymbol{\mu}_{k,b} + \mathbf{B}_k \mathbf{E}[\boldsymbol{\eta}_{k,t} | \boldsymbol{y}_t] \tag{48}$$

We consider the noise variability $\boldsymbol{n}_{k,t}$ undesired and want to remove it from $\boldsymbol{y}_t$ to predict the underlying clean speech

$$\hat{\boldsymbol{x}}_t = \sum_{k=1}^K p(k|\boldsymbol{y}_t)(\boldsymbol{y}_t - \boldsymbol{n}_{k,t}) = \boldsymbol{y}_t - \sum_{k=1}^K p(k|\boldsymbol{y}_t) \boldsymbol{n}_{k,t} \tag{49}$$

To compute $\mathbf{E}[\boldsymbol{\eta}_{k,t}|\boldsymbol{y}_t]$, one can first compute

$$\mathbf{E}[\boldsymbol{\zeta}_{k,t}|\boldsymbol{y}_t] = \mathbf{G}_k \mathbf{W}_k^{\mathsf{T}} \boldsymbol{\Psi}_k^{-1} (\boldsymbol{y}_t - \boldsymbol{\mu}_k) \tag{50}$$

with $\mathbf{G}_k = (\mathbf{I}+\mathbf{W}_k^{\mathsf{T}}\mathbf{\Psi}_k^{-1}\mathbf{W}_k)^{-1}$ and then take $\mathbf{E}[\boldsymbol{\eta}_{k,t}|\boldsymbol{y}_t]$ as a sub-vector of $\mathbf{E}[\boldsymbol{\zeta}_{k,t}|\boldsymbol{y}_t]$. Or one can explicitly write $\mathbf{E}[\boldsymbol{\eta}_{k,t}|\boldsymbol{y}_t]$ in a sub-matrix expression for computational advantages.

Eq.49 shows that the mixture of joint factor analyzers investigated here is equivalent to a piece-wise linear mapping function where each joint factor analyzer contributes locally as a linear mapping function from its region in the input feature space defined by the soft cluster of the mixture model.

5. Experimental Results

Experiments were conducted on English LVCSR. The clean speech has 60 hours of data. The noisy speech is multi-conditional, artificially generated by corrupting another set of clean speech with 10 types of background noise including **M109**, **Buccaneer**, **Leopard**, **wheel carrier**, **destroyer operation room**, **HF radio**, **babble**, **factory**, **car** and **white noise**. The SNR of each utterance in the noisy channel is randomly chosen from the 25dB to 10dB range. There is no speaker overlap between the clean and noisy speech.

The test sets consist of two scenarios: unseen conditions (Set A) and real conditions (Set B). Set A is composed of noisy speech (4 speakers, 1.1 hours) artificially generated from 5 types of background noise that didn't appear in the training set. They are **Lynx**, **machine gun**, **STITEL**, **F-16** and **pink noise**. The SNR of each utterance is randomly chosen from the range of 15dB to 5dB. The Set B (7 speakers, 1.9 hours) consists of speech recorded in real-world humvee-tank noise with SNRs estimated at 5-8dB. All the noise samples are from the NOISEX-92 dataset [11].

The feature space is constructed by splicing 9 frames of 24-dim PLP features and then projecting down to a 40-dim linear discriminant analysis (LDA) space with a global semi-tied covariance (STC) transformation. Both ML models and discriminative (FMMI and BMMI [10]) models are trained. The speaker factor dictionaries $\mathbf{A}_k$ of the mixture of Bayesian joint factor analyzers are initialized by PPCA and iteratively updated for 30 iterations. The noise factor dictionaries $\mathbf{B}_k$ are initialized by PPCA and iteratively updated for 20 iterations. There are 512 components in the mixture. Fig.1 demonstrates the sparsity of the speaker (upper panel) and noise (lower panel) factor dictionaries from one of the 512 components. The inverse of the hyper-parameter α is shown in the figure. For those α's that are driven to infinity, the corresponding columns (or atoms) in the factor dictionary are switched off.

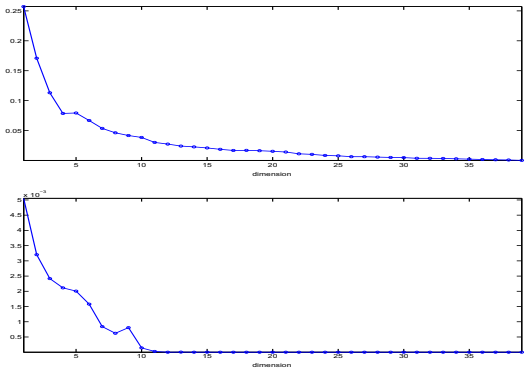

Figure 1: Sparsity of columns in speaker (upper panel) and noise (lower panel) factor dictionaries. $\frac{1}{\alpha}$ is shown in the figure.

Two sets of experiments were conducted. One is decoding compensated noisy speech with a clean acoustic model. The other is decoding compensated noisy speech with a multi-style (MST) acoustic model trained on compensated noisy speech.

The clean acoustic model consisting of 5K quinphone states and 100K Gaussians is trained on 280 hours of clean speech. The trigram language model has 330K n-grams built on a vocabulary of 45K words with 56K pronunciations. Table 1 shows the baseline performance and performance with the proposed JFA compensation. If the JFA compensation is performed in the LDA feature space, the WER improves from 33.6% to 33.2% for Set A and 36.2% to 22.3% for Set B. If the JFA compensation is performed in the FMMI feature space, the WER improves from 26.5% to 26.1% for Set A and 22.2% to 21.5% for Set B. Feature space maximum likelihood linear regression (FMLLR) is applied in all experiments.

dataset	Set A	Set B
LDA+FMLLR	33.6	36.2
FMMI+BMMI+FMLLR	26.5	22.2
LDA+JFA+FMLLR	33.2	22.3
FMMI+JFA+BMMI+FMLLR	26.1	21.5

Table 1: WERs(%) of baseline and JFA compensation on clean acoustic models.

Table 2 shows the performance on MST models. The MST models have 2K quinphone states and 80K Gaussians. The ML and FMMI+BMMI baseline models shown in the first two rows of the table are trained on the 60 hours of noisy speech. The next two rows are the WERs of JFA compensated MST models where the JFA is first estimated from the 60 hours of clean speech and 60 hours of noisy speech, and then the estimated JFA is used to compensate the 60 hours of noisy speech which is then re-trained for the MST model. The LM and dictionary are the same as the clean acoustic model decoding. Again, FMLLR is applied to all experiments. JFA compensation is performed in the LDA space. For ML models, the JFA compensation improves WERs from 43.1% to 36.7% for Set A and 20.6% to 19.3% for Set B. When FMMI+BMMI models are further trained, the improvements are from 39.1% to 32.3% for Set A and 19.0% to 17.3% for Set B.

dataset	Set A	Set B
LDA+FMLLR	43.1	20.6
FMMI+BMMI+FMLLR	39.1	19.0
LDA+JFA+FMLLR	36.7	19.3
LDA+JFA+FMMI+BMMI+FMLLR	32.3	17.3

Table 2: WERs(%) of baseline and JFA compensation on MST acoustic models.

6. Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

7. References

[1] Kenny, P., Boulianne, G., Ouellet, P. and Dumouchel, P., "Joint Factor Analysis versus Eigenchannels in Speaker Recognition", IEEE Trans. Audio, Speech and Language Proc., 15(4):1435-1447, 2007.

[2] Kenny, P., Boulianne, G., Ouellet, P. and Dumouchel, P., "Speaker and Session Variability in GMM-Based Speaker Verification", IEEE Trans. Audio, Speech and Language Proc., 15(4):1448–1460, 2007.

[3] Kenny, P., Ouellet, Dehak, N., Gupta, V. and and Dumouchel, P., "A Study of Interspeaker Variability in Speaker Verification", IEEE Trans. Audio, Speech and Language Proc., 16(5):980–988, 2008.

[4] Kenny, P., Dehak, R., Dumouchel, P. and Ouellet, P., "Front-End Factor Analysis for Speaker Verification", IEEE Trans. Audio, Speech and Language Proc., 19(4):788–798, 2010.

[5] Tipping, M. E.,"Sparse Bayesian Learning and the Relevance Vector Machine", Journal of Machine Learning Research, 1:211–244, 2001.

[6] Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society, Series B, 39(1):1–38, 1977.

[7] Bishop, C. M., "Pattern recognition and machine learning", Springer, 2006.

[8] Tipping, M. E. and Bishop, C. M.,"Mixtures of probabilistic principal component analysers", Neural Computation, 11(2):443–482, MIT Press, 1999.

[9] Cui, X., Afify, M., Saon, G. and Geol, V.,"Sparse Bayesian Factor Analysis for Stereo-based Stochastic Mapping", Interspeech,2012.

[10] Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G. and Visweswariah K., "Boosted MMI for Model and Feature-Space Discriminative Training", Proc. of ICASSP, pp. 4057–4060, 2008.

[11] Varga, A. P., Steeneken, H. J. M., Tomlinson, M. and Jones,D., "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition", Tech. Rep., DRA Speech Res. Unit, 1992.

[12] Gopinath, R. A., Ramabhadran, B., Dharanipragada, S., "Factor analysis invariant to linear transformations of data", ICASSP 1998.

[13] Ghahramani, Z., Hinton, G. E., "The EM Algorithm for Mixtures of Factor Analyzers", TR-96-1, University of Toronto.

[14] Liao, H., Gales, M. J. F., "Joint Uncertainty Decoding for Noise Robust Speech Recognition", Interspeech 2005.

[15] Kim, D. K., Gales, M. J. F., "Adaptive Training with Noisy Constrained Maximum Likelihood Linear Regression for Noise Robust Speech Recognition", Interspeech 2009.

[16] Deng, L., Droppo, J., Acero, A., "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition", IEEE Trans. Speech and Audio Proc., 11(6):568-580, 2003.

[17] Deng, L., Acero, A., Plumpe, M., Huang, X., "Large-Vocabulary Speech Recognition under Adverse Acoustic Environments", ICSLP 2000.