VARIATIONAL BAYESIAN JOINT FACTOR ANALYSIS FOR SPEAKER VERIFICATION

Xianyu Zhao[1], Yuan Dong[1,2], Jian Zhao[2], Liang Lu[2], Jiqing Liu[2], Haila Wang[1]

[1]France Telecom R&D Center (Beijing), Beijing, 100080, P.R.China
[2]Beijing University of Posts and Telecommunications, Beijing, 100876, P.R.China
{xianyu.zhao, yuan.dong, haila.wang}@orange-ftgroup.com

ABSTRACT

Joint factor analysis (JFA) has been successfully applied to speaker verification tasks to tackle speaker and session variability. In the sense of Bayesian statistics, it is beneficial to take account of the uncertainties in JFA to better characterize its speaker enrollment and verification processes, e.g. representing target speaker model by posteriori distribution of latent speaker factors and evaluating model likelihood by integrating over all latent factors. However, in a JFA model which has a large number of latent factors, it is computationally demanding to carry out these things in their exact form. In this paper, an alternative approach based on variational Bayesian is developed to explore uncertainties in JFA in an approximate yet efficient way. In this method, fully correlated posteriori distribution is approximated by a variational distribution of factorial form to facilitate inference; and a tight lower bound on model likelihood is derived. Experimental results on the 10sec4w-10sec4w task of the 2006 NIST SRE show that variational Bayesian JFA could obtain better performance than JFA using point estimate.

Index Terms—speaker verification, Gaussian mixture model, joint factor analysis, Bayesian statistics, variational approximation

1. INTRODUCTION

In recent years, joint factor analysis (JFA) has been successfully applied in many text independent speaker verification systems to deal with speaker and session variability in Gaussian mixtures models (GMMs) [1-5]. In the general Bayesian framework of JFA [1, 2], speaker and channel factors are introduced as latent or hidden random variables which have a proper priori distribution to characterize priori knowledge about speaker and session variability; target speaker model is represented by the posteriori distribution of latent speaker factors on enrollment data and likelihood score of JFA model is evaluated by integrating over all unobserved latent variables. This Bayesian framework could effectively account for uncertainties in the model and provide a better characterization of the speaker verification process.

However, in a JFA model which has a large number of latent variables, it is computationally expensive to carry out Bayesian inference in its exact form as all of the latent variables become correlated with each other in the posteriori distribution. This leads to using maximum likelihood (ML) or maximum a posteriori (MAP) point estimate of latent factors in many implementations of JFA [3-5]. The problem with point estimate lies in that it might not be reliable with limited data. Some experimental results also suggested that ignoring the uncertainties in JFA would degrade speaker verification performance [2].

To explore the uncertainties in JFA while reducing computational complexity of fully Bayesian inference, variational Bayesian JFA was proposed in this paper. In this method, the exact posteriori distribution of all latent factors in JFA is approximated by a variational distribution of factorial form from which marginal posteriori distribution over latent speaker factors (as well as latent channel factors) can be derived efficiently. The factorized approximation is then optimized through variational Bayesian to minimize the Kullback-Leibler divergence between it and the true posteriori. The log likelihood of JFA model integrating over all latent factors is also approximated by a tight lower bound based on variational Bayesian. Experimental results on the 10sec4w-10sec4w task of the 2006 NIST Speaker Recognition Evaluation (SRE) show that the variational Bayesian approach can effectively account for uncertainties in JFA and obtain better performance than using only point estimate in JFA.

2. VARIATIONAL BAYESIAN JOINT FACTOR ANALYSIS FOR SPEAKER VERIFICATION

The Gaussian Mixture Model – Universal Background Model (GMM-UBM) systems [6] are widely used for text independent speaker verification tasks. In these systems, a GMM with C components is parameterized by $\lambda = \{w_c, \mu_c, \Sigma_c; c = 1, \cdots, C\}$, where w_c, μ_c and Σ_c are the component weight, mean vector and covariance matrix of the c-th component respectively. A GMM mean supervector is constructed by concatenating all of the mean vectors in corresponding GMM:

$$M = \begin{bmatrix} \mu_1^T & \mu_2^T & \cdots & \mu_C^T \end{bmatrix}^T, \quad (1)$$

whose dimension is CF where F is the dimension of acoustic feature vectors.

Joint factor analysis is a technique to model speaker and session variability in GMMs [1, 2]. It has a latent description of the form:

$$M = m + Vy + Dz + Ux. \quad (2)$$

In this model, m characterizes the mean of speaker supervectors (in our study, m is set to be the supervector of UBM). Both V and U are low rank transformation matrices. The columns in V are usually called eigenvoices and it is assumed that the majority of speaker variability is contained in the subspace spanned by the eigenvoices. And, the columns in U are referred to as eigenchannels which are used to characterize the effects of session or channel variability. D is a $CF \times CF$ diagonal matrix which provides the possibility to model residual speaker variability that is not contained in the low dimensional subspace spanned by eigenvoices. x, y and z are random vectors. In this paper, the

components of y and z are referred to generally as speaker factors; and the components of x are called channel factors.

A Gaussian priori is used for these speaker and channel factors, which assumes statistical independence among them [1, 2]:

$$\begin{aligned} P(H) &= P_1(y)P_2(z)P_3(x) \\ &= N(y;\mu_y,B_y)N(z;\mu_z,B_z)N(x;\mu_x,B_x) \end{aligned} \tag{3}$$

where we refer to all latent factors in JFA by H for abbreviation. All covariance matrices in this priori distribution are restricted to be diagonal and all mean vectors are set to be zero.

In the sense of Bayesian statistics, given enrollment data of target speaker, $\mathcal{X}$ (of T acoustic feature vectors, $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_T$), we can get the posteriori distribution of these latent factors, $P(H|\mathcal{X})$. And, speaker dependent model is characterized by the marginal posteriori distribution of latent speaker factors:

$$P(y,z|\mathcal{X}) = \int P(x,y,z|\mathcal{X})dx \tag{4}$$

For Gaussian priori, the posteriori distribution is also Gaussian and can be calculated in close form [1]. However, the latent speaker factors and channel factors become correlated with each other in the posteriori distribution. When the number of latent factors is large (e.g., the JFA in our experiments had 105,048 latent variables), manipulating the fully posteriori distribution would be computationally demanding.

2.1. Variational Bayesian approximation to posteriori distribution of JFA

An alternative approach based on variational Bayesian [7] is studied here. Instead of using the fully posteriori distribution $P(H|\mathcal{X})$, it is approximated by a variational distribution of factorial form which decouples the correlation among speaker and channel factors:

$$Q(H) = Q_1(y)Q_2(z)Q_3(x) \tag{5}$$

From this factorized distribution, marginal distribution over speaker factors (as well as channel factors) can be derived efficiently.

Variational Bayesian is then used to optimize $Q(H)$ by minimizing the Kullback-Leibler (KL) divergence between it and the true posteriori [7, 8]:

$$\begin{aligned} Q^* &= \arg\min_Q KL(Q\|P) \\ &= \arg\min_Q \int_H Q(H)\log\frac{Q(H)}{P(H|\mathcal{X})}dH \end{aligned} \tag{6}$$

With Bayes theorem, the KL divergence can be written as:

$$\begin{aligned} KL(Q\|P) &= \int Q(H)\log\frac{Q(H)}{P(H,\mathcal{X})}dH + \log P(\mathcal{X}) \\ &= \log P(\mathcal{X}) - L(Q). \end{aligned} \tag{7}$$

Here, $L(Q)$ is defined to be an auxiliary function:

$$L(Q) = \langle \log P(H,\mathcal{X}) \rangle_Q + \mathbb{H}(Q) \tag{8}$$

where $\langle\cdot\rangle_Q$ stands for an expectation with respect to the distribution Q and $\mathbb{H}(Q)$ is the entropy of Q. As $\log P(\mathcal{X})$ does not depend on Q, minimizing KL divergence can be achieved by maximizing the auxiliary function:

$$Q^* = \arg\max_Q L(Q) \tag{9}$$

For the variational distribution of factorial form in (5), it can be shown that the optimal solution of each factor, Q_j^*, can be written in terms of its logarithm [8]:

$$\log(Q_j^*) = \langle \log P(H,\mathcal{X}) \rangle_{-Q_j} + const, \tag{10}$$

where the notation, $\langle\cdot\rangle_{-Q_j}$, represents an expectation with respect to all factors except Q_j. Equation (10) is an implicit solution as it depends on the settings of other factors. Hence, an iterative procedure can be applied which updates each factor in turn until $L(Q)$ converges. This is similar to the Gauss-Seidel iterative process proposed in [3, 4].

For speaker factors in y, its $Q_1^*(y)$ can then be written to be:

$$\begin{aligned} &\log Q_1^*(y) = \langle \log P(H,\mathcal{X}) \rangle_{-Q_1} + const \\ &= \log P_1(y) + \sum_t \sum_c Q_{\gamma_t}(\gamma_t = c) \langle \log P(\mathcal{X}_t|y,z,x,c) \rangle_{Q_2Q_3} + const \end{aligned} \tag{11}$$

where the random variable γ_t represents a component selector indicating which Gaussian component in GMM the t-th acoustic feature is drawn from; and Q_{γ_t} is the posteriori distribution of γ_t. For the c-th component, its zero and first order statistics can then defined to be:

$$N_c = \sum_t Q_{\gamma_t}(\gamma_t = c), \quad F_c = \sum_t Q_{\gamma_t}(\gamma_t = c)\mathcal{X}_t. \tag{12}$$

After some algebraic manipulation, from (11) we can then get that $Q_1^*(y)$ follows a Gaussian distribution with mean:

$$\mu_y' = L_y^{-1}\left[B_y^{-1}\mu_y + V^T\Sigma^{-1}\left(F - N\left(m + D\mu_z' + U\mu_x'\right)\right)\right], \tag{13}$$

and covariance matrix:

$$B_y' = L_y^{-1} \tag{14}$$

where N is a $CF \times CF$ diagonal matrix whose diagonal blocks are $N_c I$ (c=1, 2…C), F is a $CF \times 1$ supervector by stacking F_c (c=1, 2…C), Σ is $CF \times CF$ block diagonal matrix whose diagonal blocks are Σ_c (c=1, 2…C) and L_y is set to be $\left(B_y^{-1} + V^T N\Sigma^{-1}V\right)$; μ_z' and μ_x' are the posteriori mean vectors of z and x with respect to current setting of $Q_2(z)$ and $Q_3(x)$.

With a similar procedure, we can derive that Q_2^* and Q_3^* are also Gaussian. Their mean vectors and covariance matrices have similar expressions as (13) and (14).

2.2. Variational Bayesian lower bound on log likelihood of JFA

Based on the factorized approximation to fully posteriori distribution, in speaker dependent joint factor analysis model, the distribution over latent factors is then specified by:

$$P_S(H) = N(y;\mu_y',B_y')N(z;\mu_z',B_z')N(x;\mu_x,B_x) \tag{15}$$

where the parameters, μ_y', B_y', μ_z' and B_z', are optimized by variational Bayesian algorithm over enrollment data (as presented in Section 2.1); for the channel factors, we still use the priori distribution specified in (3) and ignore their posteriori distribution learned from enrollment data.

At verification time, given test data, $\mathcal{X}$, model likelihood is evaluated by integrating over all latent factors and detection score is:

$$S = \frac{1}{T}\log\frac{P_S(\mathcal{X})}{P(\mathcal{X})} = \frac{1}{T}\log\frac{\int P_S(H)\prod_t P\left(\mathcal{X}_t\middle|\lambda(H)\right)dH}{\int P(H)\prod_t P\left(\mathcal{X}_t\middle|\lambda(H)\right)dH}. \quad (16)$$

where $P\left(\mathcal{X}_t\middle|\lambda(H)\right)$ is the GMM observation probability:

$$P\left(\mathcal{X}_t\middle|\lambda(H)\right) = \sum_{c=1}^{C} w_c N\left(\mathcal{X}_t; M_c, \Sigma_c\right), \quad (17)$$

M_c is the *c*-th subvector of the GMM supervector corresponding to the *c*-th component.

Since it is difficult to evaluate the integration directly, an approximation based on variational Bayesian is presented here.

From (7), we can see that $L(Q)$ presents a lower bound on the log likelihood with the difference being the KL divergence. It then follows that if the variational distribution is optimized (with respect to test data at verification time) to be a good approximation in terms of KL divergence, then the bound, $L\left(Q^*\right)$, will be tight and will provide a good approximation to the log likelihood:

$$\log P(\mathcal{X}) \approx L\left(Q^*\right) \quad (18)$$

For JFA, the lower bound can then be evaluated as:

$$\begin{aligned} L\left(Q^*\right) = &\sum_t\sum_c Q^*_{\gamma_t}(\gamma_t = c)\left\langle \log P\left(\mathcal{X}_t\middle|y,z,x,c\right)\right\rangle_{Q_1^*Q_2^*Q_3^*} \\ &-KL\left(Q_1^*(y)\middle\|P_1(y)\right) - KL\left(Q_2^*(z)\middle\|P_2(z)\right) \\ &-KL\left(Q_3^*(x)\middle\|P_3(x)\right) - \sum_t KL\left(Q^*_{\gamma_t}(\gamma_t)\middle\|P(\gamma_t)\right) \end{aligned} \quad (19)$$

where $P_\bullet(\bullet)$ and $Q^*_\bullet(\bullet)$ represents respectively the priori and optimized variational posteriori distributions of latent variables.

The priori distribution of component selector, γ_t, is determined by the mixture weights in UBM, i.e.:

$$P(\gamma_t = c) = w_c, \quad c = 1,2,\ldots,C. \quad (20)$$

The posteriori distributions of component selectors can also be approximated optimally through variational Bayesian. We have:

$$\log Q^*_{\gamma_t}(\gamma_t = c) = \log w_c + \left\langle \log P\left(\mathcal{X}_t\middle|y,z,x,c\right)\right\rangle_{Q_1^*Q_2^*Q_3^*} + E_t, \quad (21)$$

where E_t is a normalization quantity:

$$E_t = -\log\left(\sum_c\left(w_c \exp\left\langle \log P\left(\mathcal{X}_t\middle|y,z,x,c\right)\right\rangle_{Q_1^*Q_2^*Q_3^*}\right)\right). \quad (22)$$

Substituting (21) into (19), we get:

$$\begin{aligned} L\left(Q^*\right) = &\sum_t \log\left(\sum_c\left(w_c \exp\left\langle \log P\left(\mathcal{X}_t\middle|y,z,x,c\right)\right\rangle_{Q_1^*Q_2^*Q_3^*}\right)\right) \\ &-KL\left(Q_1^*\middle\|P_1\right) - KL\left(Q_2^*\middle\|P_2\right) - KL\left(Q_3^*\middle\|P_3\right). \end{aligned} \quad (23)$$

The expectation of observation probability with respect to the posteriori distribution of latent factors can be evaluated as:

$$\begin{aligned} &\left\langle \log P\left(\mathcal{X}_t\middle|y,z,x,c\right)\right\rangle_{Q_1^*Q_2^*Q_3^*} = \log\frac{1}{(2\pi)^{F/2}\left|\Sigma_c\right|^{1/2}} \\ &-\frac{1}{2}\left\{\left(\mathcal{X}_t - \langle M_c\rangle\right)^T\Sigma_c^{-1}\left(\mathcal{X}_t - \langle M_c\rangle\right) + \mathrm{tr}\left(V_c^T\Sigma_c^{-1}V_c\cdot Cov_{Q_1^*}(y)\right)\right. \\ &\left.+\mathrm{tr}\left(D_c^T\Sigma_c^{-1}D_c\cdot Cov_{Q_2^*}(z)\right) + \mathrm{tr}\left(U_c^T\Sigma_c^{-1}U_c\cdot Cov_{Q_3^*}(x)\right)\right\} \end{aligned} \quad (24)$$

where V_c, U_c, and D_c denote respectively the *c*-th block of V, U, and D corresponding to the *c*-th component; $Cov_{Q_j^*}(\cdot)$ represents the covariance matrix of Q_j^*. The expectation of M_c is calculated as:

$$\langle M_c\rangle = m_c + V_c\langle y\rangle_{Q_1^*} + D_c\langle z\rangle_{Q_2^*} + U_c\langle x\rangle_{Q_3^*}. \quad (25)$$

As the speaker dependent model in (15) has the same functional form as the background model in (3), we can also get a tight lower bound on $\log P_S(\mathcal{X})$:

$$\log P_S(\mathcal{X}) \approx L_S\left(Q^*\right) \quad (26)$$

The log likelihood ratio in (16) can then be approximated by these lower bounds as:

$$S \approx \left(L_S\left(Q^*\right) - L\left(Q^*\right)\right)\Big/T. \quad (27)$$

2.3. Implementation details

In our variational Bayesian JFA, Q_1, Q_2 and Q_3 are initialized to be their priori counterpart; and $Q_{\gamma_t}(\gamma_t)$ is initialized with UBM:

$$Q_{\gamma_t}(\gamma_t = c) = \frac{w_c N\left(\mathcal{X}_t; \mu_c, \Sigma_c\right)}{\sum_{m=1}^{C} w_m N\left(\mathcal{X}_t; \mu_m, \Sigma_m\right)}. \quad (28)$$

We firstly optimize the variational distributions of speaker and channel factors iteratively while fixing $Q_{\gamma_t}(\gamma_t)$; in our experiments, 5 iterations are sufficient for convergence in this stage. Then, we optimize $Q_{\gamma_t}(\gamma_t)$ to make the lower bound $L\left(Q^*\right)$ tight; in this stage, we take account of the uncertainties in speaker and channel factors. These two stages can be carried out iteratively. In our experiments, a single iteration was performed.

When evaluating log likelihood ratio between speaker dependent and background JFA models at verification time, the covariance matrices of speaker factors in these JFA models are properly scaled to constrain the variability of these factors confronting test data:

$$B_y \leftarrow \alpha\cdot B_y,\ B_z \leftarrow \alpha\cdot B_z. \quad (29)$$

We found such kind of scaling could make detection scores more stable and improve verification performance; α was set to be 0.1 in our experiments.

3. EXPERIMENTAL RESULTS

Speaker verification experiments were carried out on the 10sec4w-10sec4w task of the 2006 NIST SRE [9]. In this task, the length of enrollment and test utterances is about 10 seconds; and there are a total of 2,942 true trials and 29,608 false trials.

In our systems, the acoustic features were a 51-dimentional PLP vector. We used a gender independent UBM with 2048 Gaussians trained from the Switchboard corpora (I, II and Cellular parts). Our factor analysis model had 500 eigenvoices and 100 eigenchannels, which were estimated using PCA on the 2004 and 2005 NIST SRE corpora. As in [4], the D matrix in our JFA model was set according to the following equation, $I = \tau D^T\Sigma^{-1}D$, where τ is 16 in our experiments.

Four different schemes of JFA were compared:

- In the "Baseline" system, target speaker model is constructed by the MAP point estimate of speaker factors on enrollment data. At verification time, session factors are not used. Log likelihood ratio test is carried out through:

$$S_B = \frac{1}{T}\log\frac{\prod_t p\left(\mathcal{X}_t \middle| m + V\mu'_y + D\mu'_z\right)}{\prod_t p\left(\mathcal{X}_t \middle| m\right)} \tag{30}$$

- In the "Eigenchannel Adaptation" system, target speaker models are constructed in the same way as the "Baseline". MAP point estimates of channel factors are obtained at verification time for target speaker and background models, i.e., μ'_x and $\underline{\mu}'_x$. And likelihood ratio score is evaluated as:

$$S_A = \frac{1}{T}\log\frac{\prod_t p\left(\mathcal{X}_t \middle| m + V\mu'_y + D\mu'_z + U\mu'_x\right)}{\prod_t p\left(\mathcal{X}_t \middle| m + U\underline{\mu}'_x\right)} \tag{31}$$

This scoring scheme is similar to that used in [5].

- In the "Eigenchannel Integration" system, for speaker factors, we still use its MAP point estimate derived from enrollment data; at verification time, log-likelihood score is evaluate by integrating over channel factors to take account of uncertainty in them:

$$S_I = \frac{1}{T}\log\frac{\int P(x)\prod_t p\left(\mathcal{X}_t \middle| m + V\mu'_y + D\mu'_z + Ux\right)dx}{\int P(x)\prod_t p\left(\mathcal{X}_t \middle| m + Ux\right)dx} \tag{32}$$

An approximation to the integrations in (32) can be obtained by dropping off KL divergence and covariance terms related to speaker factors in (23) - (25).

- The "Variational Bayesian JFA" system is what we developed in this paper.

In Fig. 1, we show verification performance of these systems. Comparing "Baseline" and "Eigenchannel Adaptation", we see that incorporating channel factors for session variability can improve verification performance over the baseline. As the test utterances are very short in this task, the point estimate of channel factor might not be reliable. After integrating over the uncertainty in channel factors, better performance was obtained by "Eigenchannel Integration" than "Eigenchannel Adaptation". In the "Variational Bayesian JFA" system, we not only consider uncertainty in channel factors but also take into account the uncertainty in speaker factors due to limited enrollment data. This is proven helpful and the "Variational Bayesian JFA" system obtained the best performance among these four systems.

4. CONCLUSIONS

In this paper, variational Bayesian algorithm is used to do approximate inference in joint factor analysis for speaker verification. After approximating the fully correlated posteriori distribution of all latent factors by a variational distribution of factorial form, posteriori distribution of speaker factors can be derived efficiently without computationally demanding marginalization procedure. Target speaker models are then represented by such kind of posteriori distribution, which can better characterize the uncertainty in the enrollment process than mere point estimate. With variational Bayesian algorithm, we also derive a lower bound on the log likelihood of joint factor analysis which integrates over all latent factors to take account of uncertainties in them. Experimental results show that variational

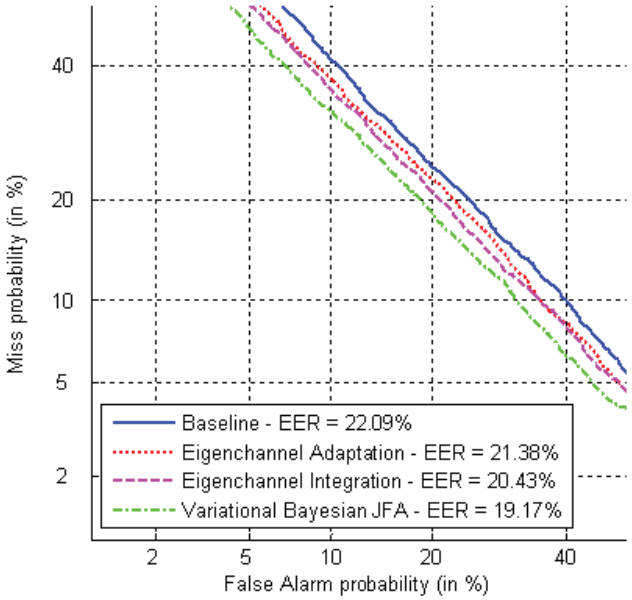

Fig.1 DET curves for different schemes of JFA on the 10sec4w-10sec4w task of the 2006 NIST SRE

Bayesian provides an effective approach to explore uncertainties in joint factor analysis and obtains promising verification performance.

5. REFERENCES

[1] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and Algorithms, Tech Report CRIM-06/08-13," 2005, Online: http://www.crim.ca/perso/patrick.kenny .

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no.4, pp. 1435-1447, May, 2007.

[3] R. Vogt, B. Baker and S. Sridharan, "Modelling session variability in text-independent speaker verification," in: *Proc. INTERSPEECH 2005*, pp. 3117-3120, Lisbon, Portugal, 2005.

[4] D. Matrouf, N. Scheffer, B. Fauve and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in: *Proc. INTERSPEECH 2007*, pp. 1242-1245, Antwerp, Belgium, 2007.

[5] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no.7, pp. 1979-1986, September, 2007.

[6] D. Reynolds, T. Quatieri and R. Dunn, "Speaker Verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19-41, 2000.

[7] H. Attias, "A variational Bayesian framework for graphical models," in: *Advances in Neural Information Processing Systems*, vol. 12, pp. 209-215, Cambridge, MA, 2000.

[8] J. Winn and C. Bishop, "Variational message passing," *J. Machine Learning Research*, vol. 6, pp. 661-694, 2005.

[9] "The NIST 2006 speaker recognition evaluation plan," Online: http://www.nist.gov./speech/tests/spk/spk/2006 .