A study of variational method for text-independent speaker recognition

[1]*Liang He,* [1]*Yao Tian,* [1]*Yi Liu,* [2]*Fang Dong,* [1]*WeiQiang Zhang,* [1]*Jia Liu*

[1]Tsinghua National Laboratory for Information Science and Technology,
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[2]School of Information and Electrical Engineering,
Zhejiang University City College, Hangzhou 310015, China

heliang,wqzhang,liuj@tsinghua.edu.cn, chinaty188@163.com
liu-yi15@mails.tsinghua.edu.cn, dongf@zucc.edu.cn

Abstract

An i-vector has become the state-of-the-art algorithm for text-independent recognition. Most of related works take the extraction of the i-vector as a black-box by using some open software (e.g. Kaldi, Alize) and focus on the vector-based back-end algorithms, such as length normalization, WCCN, or PLDA. In this paper, we study the variational method and present a concise derivation for the i-vector. Based on our proposed methods, three criteria for derivation are compared. There are maximum likelihood (ML), maximum a posteriori (MAP) and maximum marginal likelihood (MML) criterion respectively. Experimental results on the NIST SRE08 tel-tel-English condition task proved our works.

Index Terms: Gaussian mixture models, variational method, i-vector, text-independent speaker recognition

1. Introduction

The i-vector firstly proposed by Najim Dehak has become the state-of-the-art algorithm for text-independent speaker recognition [1]. In Dehak's classic work [1], he combines the eigenvoice and eigenchannel subspaces together to form a total variability subspace. The corresponding subspace loading factor is termed as the identity vector (i-vector for short). Unfortunately, the theoretic part is not fully addressed in that paper and mainly from the joint factor analysis (JFA) proposed by Patrick J. Kenny [2, 3, 4, 5, 6]. Among Kenny's JFA related works, the core derivation is built on the *Lemma 1* in [3], *Theorem 1* in [5] and *Equation (5)* in [6]. According to [3], for example, the log conditional likelihood of $\boldsymbol{o}$ given $\boldsymbol{w}$ and the parameter set (T, Σ) is $\log p(\boldsymbol{o}|\boldsymbol{w}) = G + H(\boldsymbol{w})$, where the definition of G and $H(\boldsymbol{w})$ can be found in [3]. [1] However, if we bring GMMs with subspace model $(\sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_m + T_m\boldsymbol{w}, \Sigma_m))$ into the above mentioned log conditional likelihood, we won't get the above equation, which motivates us to re-consider the derivation of i-vector (and JFA).

We first give our log conditional likelihood (log likelihood, for simplicity) which can be written as a form of $(\log \sum \exp)$. It's non-convex and difficult to be optimized. To handle with it, we propose a concise derivation based on the variational method [7, 8]. Our proposed variational method avoids solving the log likelihood directly and tries to maximize its lower bound by the Jensen's inequality [9]. The lower bound has a form of $(\sum)$, which is convex and easy to be solved.

[1]Different from [3], we use the symbol $\boldsymbol{o}$, $\boldsymbol{w}$ and T instead of the symbol $\mathcal{X}$, $\boldsymbol{y}$ and V respectively in order to ensure that the symbols in our paper are consistent.

Our above work is carried out in the context of maximum likelihood criterion. We also extend it to the maximum a posteriori (MAP) and maximum marginal likelihood (MML) criterion [10, 11]. The MAP criterion takes the prior distribution into account which may improve the recognition performance. However, the prior selection is an open yet tough question which exists in many Bayesian estimation. Here, we examine and compare the priors from several considerations, such as the principle of maximum entropy [13], empirical Bayes method [14]. The MML criterion takes the uncertainty of model's parameters into account by integrating them out [10, 11].

The remainder is as follows: Section 2 introduces an inequality for the $(\log \sum \exp)$ operator based on the Jensen's inequality. Section 3 presents our model. Section 4 illustrates and compares variational methods based on the ML, MAP and MML criteria. Finally, experimental results and conclusion are briefly given in section 5 and 6.

2. An inequality for the $(\log \sum \exp)$ operator

Let p and q are mixture models for probability density function, $p = \sum_{m=1}^{M} \alpha_m p_m$ and $q = \sum_{m=1}^{M} \beta_m q_m$, $\forall x, m$, $p_m(x) = p_m > 0$, $\int p_m dx = 1$, $q_m(x) = q_m > 0$, $\int q_m dx = 1$, $\alpha_m > 0$, $\sum_{m=1}^{M} \alpha_m = 1$, $\beta_m > 0$, $\sum_{m=1}^{M} \beta_m = 1$. We examine the $\log(p)$

$$
\begin{aligned}
\log(p) &= \log\big(\sum_{m=1}^{M} \alpha_m p_m\big) \\
&= \log\big(\sum_{m=1}^{M} (\alpha_m p_m \frac{\beta_m q_m}{\beta_m q_m})\big) + \log(q) - \log(q) \\
&= \log\big(\sum_{m=1}^{M} \frac{\beta_m q_m}{q} \frac{\alpha_m p_m}{\beta_m q_m}\big) + \log(q) \\
&\geq \sum_{m=1}^{M} \frac{\beta_m q_m}{q} \log\big(\frac{\alpha_m p_m}{\beta_m q_m}\big) + \log(q) \\
&\geq \sum_{m=1}^{M} \frac{\beta_m q_m}{q} \log\big(\alpha_m p_m\big) + \log(q) - \sum_{m=1}^{M} \frac{\beta_m q_m}{q} \log(\beta_m q_m)
\end{aligned} \tag{1}
$$

Note that $\sum_{m=1}^{M} (\beta_m q_m)/q = 1$ and the above inequality is from the Jensen's inequality [12]. The p is the probability density function we interested but complex. In most cases, there is no analytical solution for the maximizing $\log p$ problem. From the

perspective of variational method, we avoid solving the objective function directly and turn to maximizing the lower bound by selecting a proper q with a simple structure or known parameters. Given a proper q, the above inequality's last two items become constants and $\beta_m q_m/q$ is easier to be obtained. Our maximizing problem reduces to linear weighted summation of $\log(\alpha_m p_m)$. This is especially suitable for mixture models with exponential family distributions which are widely used in the field of machine learning. It transforms a complex $(\log \sum \exp)$ problem into a simple $(\sum)$ problem.

Further more, we re-consider it from the view of information theory by re-arranging and integrating

$$\begin{aligned} D_{\text{KL}}(q\|p) &= \int q \log(\frac{q}{p}) dx \\ &\leq \int \left(\sum_{m=1}^{M} \beta_m q_m \log(\frac{q_m}{p_m}) + \sum_{m=1}^{M} q_m \beta_m \log(\frac{\beta_m}{\alpha_m}) \right) dx \\ &\leq \sum_{m=1}^{M} \beta_m D_{\text{KL}}(q_m\|p_m) + D_{\text{KL}}(\beta\|\alpha) \end{aligned} \tag{2}$$

This inequality states that the KL divergence between two mixture models is upper bounded by two types of divergences: a weighted summation of mixture component divergences and weight divergence. These two items are loosely coupled, which means an iterative algorithm for the minimization of the upper bound of KL divergence, see Algorithm 1. In some cases, $D_{\text{KL}}(\beta\|\alpha)$ is trivial and we can simplify the objective function to $\min \sum_{m=1}^{M} \beta_m D_{\text{KL}}(q_m\|p_m)$ and speed up the optimization.

Algorithm 1 Finding a q with a minimal $D_{\text{KL}}(q\|p)$ for an unknown p

while a stopping criterion is not met **do**
1. select a mixture model q by which we compute β easily for each x
2. minimize a weighted sum of mixture divergences with computed β
3. update β by the q_m obtained in the previous step

end while

3. Model assumption

The high degree of freedom of GMM and the limited duration of utterances for enrolling a speaker or testing become a dilemma. The subspace method which has low degree of freedom addresses this problem. In our work, the model assumption is the same to the i-vector [1].

$$\boldsymbol{\mu} = \boldsymbol{\mu}_{\text{ubm}} + T\boldsymbol{w} \tag{3}$$

Let a GMM λ_G with its weights ω_m, means $\boldsymbol{\mu}_m$ and covariance Σ_m

$$p(\boldsymbol{o}|\lambda_G) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_m, \Sigma_m) \tag{4}$$

The subspace method which constrains the mean parameters to lie in a linear subspace is

$$p(\boldsymbol{o}|\lambda_S) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_m + T_m\boldsymbol{w}, \Sigma_m) \tag{5}$$

The $p(\boldsymbol{o}|\lambda_S)$ is non-convex and the variational method offers a concise solution to the intractable problem by maximizing the lower bound.

4. Variational method

4.1. Maximum likelihood criterion

To obtain a proper T and Σ, our goal is to maximize a log likelihood function over a database $\mathcal{O}$

$$\begin{aligned} &\arg\max_{T,\Sigma} \log p(\mathcal{O}|\lambda_S) \\ =&\arg\max_{T,\Sigma} \sum_{n=1}^{N} \log p(O_n|\lambda_S) \\ =&\arg\max_{T,\Sigma} \sum_{n=1}^{N} \sum_{k_n=1}^{K_n} \log p(\boldsymbol{o}_{k_n}|\lambda_S) \\ =&\arg\max_{T,\Sigma} \sum_{n=1}^{N} \sum_{k_n=1}^{K_n} \log \left(\sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{o}_{k_n}|\boldsymbol{\mu}_m + T_m\boldsymbol{w}_n, \Sigma_m) \right) \end{aligned} \tag{6}$$

where K_n is the duration of the n-th utterance. Note that, this log likelihood function is non-convex and we try to maximize its lower bound.

We select

$$p_m = \omega_m \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_m + T_m\boldsymbol{w}_n, \Sigma_m) \tag{7}$$

with an unknown parameter set $\{\omega_m, \boldsymbol{\mu}_m, \Sigma_m, T_m, \boldsymbol{w}_n\}$ and

$$q_m = \omega_{\text{ubm},m} \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_{\text{ubm},m}, \Sigma_{\text{ubm},m}) \tag{8}$$

with a known parameter set $\{\omega_{\text{ubm},m}, \boldsymbol{\mu}_{\text{ubm},m}, \Sigma_{\text{ubm},m}, T_m, \boldsymbol{w}_n = \boldsymbol{0}\}$.

So, by the inequality (1) we have

$$\begin{aligned} &\log \left(\sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{o}_{k_n}|\boldsymbol{\mu}_m + T_m\boldsymbol{w}, \Sigma_m) \right) \geq \\ &\sum_{m=1}^{M} \gamma_{\text{ubm},m}(\boldsymbol{o}_{k_n}) \Bigg(\log \omega_m - \frac{F}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_m| \\ &- \frac{1}{2} \left(\boldsymbol{o}_{k_n} - \boldsymbol{\mu}_{\text{ubm},m} - T_m\boldsymbol{w}_n\right)^t \Sigma_m^{-1} \left(\boldsymbol{o}_{k_n} - \boldsymbol{\mu}_{\text{ubm},m} - T_m\boldsymbol{w}_n\right) \Bigg) \\ &+ \text{const} \end{aligned} \tag{9}$$

where $\gamma_{\text{ubm},m}(\boldsymbol{o}_{k_n})$ is as follows

$$\gamma_{\text{ubm},m}(\boldsymbol{o}_{k_n}) = \frac{\omega_m \mathcal{N}(\boldsymbol{o}_{k_n}|\boldsymbol{\mu}_m + T_m\boldsymbol{w}_n, \Sigma_m)}{\sum_{m'=1}^{M} \omega_{m'} \mathcal{N}(\boldsymbol{o}_{k_n}|\boldsymbol{\mu}_{m'} + T_{m'}\boldsymbol{w}_n, \Sigma_{m'})} \tag{10}$$

Thus, the objective function is lower bounded by the one column equation (1), and

$$\begin{aligned} Z_{n,m} &= \sum_{k_n=1}^{K_n} \gamma_{\text{ubm},m}(\boldsymbol{o}_{k_n}) \\ F_{n,m} &= \sum_{k_n=1}^{K_n} \gamma_{\text{ubm},m}(\boldsymbol{o}_{k_n})(\boldsymbol{o}_{k_n} - \boldsymbol{\mu}_{\text{ubm},m}) \\ S_{n,m} &= \text{diag} \left\{ \sum_{k_n=1}^{K_n} \gamma_{\text{ubm},m}(\boldsymbol{o}_{k_n})(\boldsymbol{o}_t - \boldsymbol{\mu}_{\text{ubm},m})(\boldsymbol{o}_{k_n} - \boldsymbol{\mu}_{\text{ubm},m})^t \right\} \end{aligned} \tag{11}$$

The above equation is a quadratic programming and we solve it by an two-step iterative algorithm. In the first step, $\{T, \Sigma\}$ are fixed. The optimized $\boldsymbol{w}_n$ for a given O_n is obtained by the derivation of $\boldsymbol{w}_n$,

$$\boldsymbol{w}_n = \left(T^t Z_n \Sigma^{-1} T\right)^{-1} T^t \Sigma^{-1} F_n \tag{12}$$

$$
\begin{aligned}
&\sum_{n=1}^{N}\sum_{k_n=1}^{K_n}\log\left(\sum_{m=1}^{M}\omega_m\mathcal{N}(\boldsymbol{o}_{k_n}|\boldsymbol{\mu}_m+T_m\boldsymbol{w}_n,\Sigma_m)\right)\\
&\geq\sum_{n=1}^{N}\sum_{k_n=1}^{K_n}\sum_{m=1}^{M}\gamma_{\text{ubm},m}(\boldsymbol{o}_{k_n})\Big(\log\omega_m+\log\mathcal{N}(\boldsymbol{o}_{k_n}|\boldsymbol{\mu}_m+T_m\boldsymbol{w},\Sigma_m)\Big)+\text{const} \qquad (1)\\
&\geq\sum_{n=1}^{N}\sum_{m=1}^{M}\left(Z_{n,m}\log\frac{\omega_m}{(2\pi)^{\frac{F}{2}}\log|\Sigma_m|^{\frac{1}{2}}}-\frac{1}{2}\text{tr}(S_{n,m}\Sigma_m^{-1})+F_{n,m}^t\Sigma_m^{-1}T_m\boldsymbol{w}_n-\frac{1}{2}\boldsymbol{w}_n^k\left(T_m^tZ_{n,m}\Sigma_m^{-1}T_m\right)\boldsymbol{w}_n\right)+\text{const}
\end{aligned}
$$

In the second step, $\boldsymbol{w}_n$ is already estimated for each training utterance. The optimized T_i is obtained by the derivation of T_i

$$\sum_{n=1}^{N}F_{n,m}\boldsymbol{w}_n^t=\sum_{n=1}^{N}Z_nT_m\boldsymbol{w}_n\boldsymbol{w}_n^t \qquad (13)$$

and the optimized Σ_i is obtained by the derivation of Σ_i

$$
\begin{aligned}
\sum_{n=1}^{N}Z_{n,m}\Sigma_m=\text{diag}\Big\{\sum_{n=1}^{N}\big[&S_{n,m}-F_{n,m}(T_m\boldsymbol{w}_n)^t\\
&-(T_m\boldsymbol{w}_n)F_{n,m}^t+(T_m\boldsymbol{w}_n)(T_m\boldsymbol{w}_n)^t\big]\Big\}
\end{aligned} \qquad (14)
$$

After a few iterations ($6\sim20$ times), we get reasonable T and Σ.

4.2. Maximum a posteriori criterion

Another way to obtain proper T and Σ is by maximizing the posterior probability $p(\lambda|\mathcal{O})$. According to Bayes' rules, the objective function is formulated as

$$p(\lambda|\mathcal{O})=\frac{p(\mathcal{O}|\lambda)p(\lambda)}{p(\mathcal{O})}\propto\prod_{n=1}^{N}p(O_n|\lambda_n)p(\lambda_n) \qquad (15)$$

The prior distribution can be selected from different considerations. In this section, we assume the prior $p(\lambda_n)$ as a Gaussian distribution and have

$$
\begin{aligned}
&\arg\max_{T,\Sigma}\ \log p(\lambda_S|\mathcal{O})\\
=&\arg\max_{T,\Sigma}\sum_{n=1}^{N}\left\{\left[\sum_{k_n=1}^{K_n}\log\left(\sum_{m=1}^{M}f(\boldsymbol{o}_{k_n}|\lambda_m)\right)\right]\right.\\
&\left.+\log\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_w,\Sigma_w)\right\}
\end{aligned} \qquad (16)
$$

By following the procedure in the above section and discarding the const items, we obtain the lower bound of $\log p(\lambda|\mathcal{O})$, see the one column equation (2). So, the estimation of $\boldsymbol{w}_n$ in the first step is as follows

$$\boldsymbol{w}_n=\left(\Sigma_w^{-1}+T^tZ_n\Sigma^{-1}T\right)^{-1}\left(\Sigma_w^{-1}\mu_w+T^t\Sigma^{-1}F_n\right) \qquad (17)$$

and the estimation of T and Σ in the second step is the same to equation (13) and (14).

4.2.1. Principle of maximum entropy, ME

From the principle of maximum entropy, we should select a prior which is the most uninformative distribution. In other words, this selection makes no propensity to any candidate model. From this consideration, $\boldsymbol{\mu}_w$ and Σ_w are zero vector and identity matrix respectively.

4.2.2. Empirical Bayes method, EBM

In the setting of empirical Bayes method, the prior distribution is estimated from the data. It is biased on the given information (from data used to estimate prior distribution) and maximally noncommittal with regard to missing information. In our case, $\boldsymbol{\mu}_w$ and Σ_w are $\boldsymbol{\mu}_w=\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{w}_n$ and $\Sigma_w=\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{w}_n-\boldsymbol{\mu}_w)(\boldsymbol{w}_n-\boldsymbol{\mu}_w)^t$ respectively from this consideration. Similar works are [15] and [16]. Both of them tried to estimate prior distribution from auxiliary database to made use of existing information.

4.3. Maximum marginal likelihood criterion

The marginal likelihood is obtained by integrating $\boldsymbol{w}$, see the one column equation (3). This equation can be rewritten as the one column equation (4). For simplicity, let

$$
\begin{aligned}
\Sigma_\alpha&=\left(T_m^t\Sigma_m^{-1}T_m+\frac{1}{K}\Sigma_w^{-1}\right)^{-1}\\
\Sigma_{\beta,m}&=(\Sigma_m^{-1}-\Sigma_m^{-1}T_m\Sigma_\alpha T_m^t\Sigma_m^{-1})^{-1}\\
T_{\beta,m}&=\frac{1}{K}\Sigma_\beta^{-1}\Sigma_m^{-1}T_m\Sigma_\alpha\Sigma_w^{-1}
\end{aligned} \qquad (18)
$$

And we get a clear version of $p(\boldsymbol{o})$

$$
\begin{aligned}
p(\boldsymbol{o})\propto\sum_{m=1}^{M}\omega_m|\Sigma_{\beta,m}|^{-\frac{1}{2}}\exp\Big\{&-\frac{1}{2}(\boldsymbol{o}_k-\boldsymbol{\mu}_{\text{u},m}-T_{\beta,m}\boldsymbol{\mu}_w)^t\\
&\Sigma_{\beta,m}^{-1}(\boldsymbol{o}_k-\boldsymbol{\mu}_{\text{u},m}-T_{\beta,m}\boldsymbol{\mu}_w)\Big\}
\end{aligned} \qquad (19)
$$

which is the same to maximum likelihood criterion in form. Hence, it means that the ML and MML criteria lead to the same result.

4.4. Discussion

The basic mathematical problem for the above three sections is the same ($(\log\sum\exp)$). The objective function is not convex and the variational method provides a flexible solution by maximizing its lower bound (inequality (1)). Our presented inequality is powerful by carefully selecting p and q. In our paper, we just give a selection of p and q, see Table 1. There are more suitable options. For example, the q or p can be selected as GMMs with full covariance matrices [17], or the q can be selected from the deep neural network [18].

5. Experiments

Experiments were carried out on the female common condition 7 of NIST SRE 2008 core task (female-c7-08). We used previous NIST SRE data corpus to train UBM, T, Σ.

$$\arg\max_{T,\Sigma} \sum_{n=1}^{N} \Bigg[\sum_{m=1}^{M} \Bigg(Z_{n,m} \log \omega_m - Z_{n,m} \frac{F}{2} \log(2\pi) - Z_{n,m} \frac{1}{2} \log |\Sigma_m| - \frac{1}{2} \mathrm{tr}(S_{n,m} \Sigma_m^{-1}) + F_{n,m}^t \Sigma_m^{-1} T_m \boldsymbol{w}_n - \frac{1}{2} \boldsymbol{w}_n^k \left(T_m^t Z_{n,m} \Sigma_m^{-1} T_m \right) \boldsymbol{w}_n \Bigg) - \frac{1}{2} (\boldsymbol{w}_n - \boldsymbol{\mu}_w)^t \Sigma_w^{-1} (\boldsymbol{w}_n - \boldsymbol{\mu}_w) \Bigg] \tag{2}$$

$$\begin{aligned} p(\mathcal{O}) &= \int p(\mathcal{O}|\boldsymbol{w}) p(\boldsymbol{w}) d\boldsymbol{w} \\ &= \int \left\{ \prod_{k=1}^{K} \left[\sum_{m=1}^{M} \omega_m (2\pi)^{-\frac{F}{2}} |\Sigma_m|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{o}_k - \boldsymbol{\mu}_{\mathrm{u},m} - T_m \boldsymbol{w})^t \Sigma_m^{-1} (\boldsymbol{o}_k - \boldsymbol{\mu}_{\mathrm{u},m} - T_m \boldsymbol{w}) \right) \right] \right\} \\ &\quad (2\pi)^{-\frac{D}{2}} |\Sigma_w|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{w} - \boldsymbol{\mu}_w)^t \Sigma_w^{-1} (\boldsymbol{w} - \boldsymbol{\mu}_w) \right) d\boldsymbol{w} \end{aligned} \tag{3}$$

$$\begin{aligned} p(\boldsymbol{o}) &= \int \Bigg[\sum_{m=1}^{M} \omega_m (2\pi)^{-\frac{F}{2} - \frac{D}{2K}} |\Sigma_m|^{-\frac{1}{2}} |\Sigma_w|^{-\frac{1}{2K}} \exp\left(-\frac{1}{2} (\boldsymbol{o}_k - \boldsymbol{\mu}_{\mathrm{u},m})^t \Sigma_m^{-1} (\boldsymbol{o}_k - \boldsymbol{\mu}_{\mathrm{u},m}) \right. \\ &\quad \left. - \frac{1}{2K} \boldsymbol{\mu}_w^t \Sigma_w^{-1} \boldsymbol{\mu}_w - \frac{1}{2} \boldsymbol{w}^t \left(T_m^t \Sigma_m^{-1} T_m + \frac{1}{K} \Sigma_w^{-1} \right) \boldsymbol{w} + \left((\boldsymbol{o}_k - \boldsymbol{\mu}_{\mathrm{u},m})^t \Sigma_m^{-1} T_m + \frac{1}{K} \boldsymbol{\mu}_w^t \Sigma_w^{-1} \right) \boldsymbol{w} \right) \Bigg] d\boldsymbol{w} \\ &= \sum_{m=1}^{M} \omega_m \exp\left\{ -\frac{1}{2} (\boldsymbol{o}_k - \boldsymbol{\mu}_{\mathrm{u},m})^t \Sigma_m^{-1} (\boldsymbol{o}_k - \boldsymbol{\mu}_{\mathrm{u},m}) - \frac{1}{2K} \boldsymbol{\mu}_w^t \Sigma_w^{-1} \boldsymbol{\mu}_w \right. \\ &\quad \left. + \frac{1}{2} \left((\boldsymbol{o}_k - \boldsymbol{\mu}_{\mathrm{u},m})^t \Sigma_m^{-1} T_m + \frac{1}{K} \boldsymbol{\mu}_w^t \Sigma_w^{-1} \right)^t \left(T_m^t \Sigma_m^{-1} T_m + \frac{1}{K} \Sigma_w^{-1} \right)^{-1} \left((\boldsymbol{o}_k - \boldsymbol{\mu}_{\mathrm{u},m})^t \Sigma_m^{-1} T_m + \frac{1}{K} \boldsymbol{\mu}_w^t \Sigma_w^{-1} \right) \right\} \\ &\quad \int (2\pi)^{-\frac{F}{2} - \frac{D}{2K}} |\Sigma_m|^{-\frac{1}{2}} |\Sigma_w|^{-\frac{1}{2K}} \\ &\quad \exp\left\{ -\frac{1}{2} \left[\boldsymbol{w} - \left(T_m^t \Sigma_m^{-1} T_m + \frac{1}{K} \Sigma_w^{-1} \right)^{-1} \left((\boldsymbol{o}_k - \boldsymbol{\mu}_{\mathrm{u},m})^t \Sigma_m^{-1} T_m + \frac{1}{K} \boldsymbol{\mu}_w^t \Sigma_w^{-1} \right) \right]^t \left(T_m^t \Sigma_m^{-1} T_m + \frac{1}{K} \Sigma_w^{-1} \right) \right. \\ &\quad \left. \left[\boldsymbol{w} - \left(T_m^t \Sigma_m^{-1} T_m + \frac{1}{K} \Sigma_w^{-1} \right)^{-1} \left((\boldsymbol{o}_k - \boldsymbol{\mu}_{\mathrm{u},m})^t \Sigma_m^{-1} T_m + \frac{1}{K} \boldsymbol{\mu}_w^t \Sigma_w^{-1} \right) \right] \right\} d\boldsymbol{w} \end{aligned} \tag{4}$$

Table 1: *Comparison of different criteria*

	ML	MAP, ME
$\boldsymbol{w}$ prior	fixed, unknown $\times$	fixed, unknown const
	MAP, EBM	MML
$\boldsymbol{w}$ prior	fixed, unknown from data	variable $\times$

Speech/silence segmentation was performed by a G.723.1 VAD detector. A 13-dimensional MFCC with Δ and $\Delta\Delta$ was extracted. 39-dimensional vectors were subjected to feature warping [19]. UBMs with 1024 Gaussian components were gender-dependent. The rank of T was 600. Length normalization were adopted [20]. Experimental results are presented in the Table 2.

Table 2: *Experimental results on the NIST SRE08 tel-tel-English female task*

Cosine kernel	EER(%)	MinDCF08
ML	5.75	0.250
MAP, ME (i-vector)	5.76	0.255
MAP, EBM	6.17	0.279
MML	5.75	0.250
PLDA	EER(%)	MinDCF08
ML	2.82	0.123
MAP, ME (i-vector)	2.84	0.125
MAP, EBM	3.03	0.132
MML	2.82	0.123

6. Analysis and Conclusions

In this paper, we propose a concise method for the derivation of i-vector (also suitable for JFA) based on the variational method. We compare several variational methods based on the ML, MAP and MML criteria. The ML and MML criteria lead to the same result in our setting. In the case of MAP criterion, the prior selection has an obvious influence. Although our experimental result (MAP, EBM) is inferior to the other criteria, we believe that it may achieve a better performance by selecting training data carefully which is also our future work.

7. Acknowledgements

Authors should ensure that their identities are not revealed in any way.

8. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

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

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

[5] P. Kenny, "Joint factor analysis of speaker and session variability : Theory and algorithms - Technical report CRIM-06/08-13," Montreal, CRIM, 2005. Online: http://www.crim.ca/perso/patrick.kenny/FAtheory.pdf, 2016.

[6] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of Telephone Conversations using Factor Analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, 2010.

[7] C. M Bishop, J.M.Winn, and D. Spiegelhalter. "VIBES: A variational inference engine for Bayesian networks." *In Advances in Neural Information Processing Systems*, 2002.

[8] W. F. Charles , J. R. Stephen "A tutorial on variational Bayesian inference," *Artificial Intelligence Review*, vol. 38, no. 2, pp. 85–95, 2012.

[9] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products, 6th edtion* San Diego, CA: Academic Press, p. 1101, 2000.

[10] wiki: Marginal likelihood, online: https://en.wikipedia.org/wiki/Marginal_likelihood

[11] O. Dikmen, C. Fevotte. "Maximum marginal likelihood estimation for nonnegative dictionary learning," in *Proc. ICASSP 2011*, Prague, Czech Republic, 2011, pp. 1992–1995.

[12] J. R. Hershey, P. A. Olsen "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. ICASSP 2007*, Honolulu, Hawaii, USA, 2007, pp. 317–320.

[13] P. Harremoes, F. Topsoe "Maximum Entropy Fundamentals," *Entropy*, vol. 3, no. 3, pp. 191–226, 2001.

[14] C.M. Bishop, *Neural networks for pattern recognition*. Oxford University Press, 2005.

[15] R. Travadi, M. V. Segbroeck and S. Narayanan "Modified prior i-Vector Estimation for Language Identification of Short Duration Utterances," in *Proc. INTERSPEECH 2014*, MAX Atria, Singapore, 2014, pp. 3037–3041.

[16] S. E. Shepstone, K. A. Lee, H. Z. Li, Z. H. Tan and S. H. Jensen "Source-specific informative prior for i-vector extraction," in *Proc. ICASSP 2015*, Brisbane, Australia, 2015, pp. 4185–4189.

[17] P. Matejka, O. Glembek, F. Castaldo and *et al*. "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proc. ICASSP 2011*, Prague, Czech, 2011, pp. 4828–4831.

[18] Y. Lei, L. Ferrer, M. McLaren, N. Scheffer, "A Deep Neural Network Speaker Verification System Targeting Microphone Speech," *Proc. INTERSPEECH 2014*, MAX Atria, Singapore, 2014, pp. 681–685.

[19] B. Xiang, U. V. Chaudhari, J. Navratil, G. N. Ramaswamy, and R. A. Gopinath, "Short-time gaussianization for robust speaker verification," in *Proc. ICASSP 2002*, Orlando, Florida, 2002, pp. 681–684.

[20] D. G. Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. INTERSPEECH 2011*, Florence, Italy, 2011, pp. 249-252.