Minimax i-vector extractor for short duration speaker verification

Ville Hautamäki[1,2*] , *You-Chi Cheng*[2], *Padmanabhan Rajan*[1], *and Chin-Hui Lee*[2]

[1]School of Computing, University of Eastern Finland, Finland
[2]ECE, Georgia Institute of Technology, USA
{villeh, paddy}@cs.uef.fi, {yccheng, chl}@ece.gatech.edu

Abstract

Total variability modeling, based on i-vector extraction of converting a variable-length sequence of feature vectors into a fixed-length i-vector, is currently an adopted parametrization technique for state of-the-art speaker verification systems. However, when the number of the feature vectors is low, uncertainty in the i-vector representation as a point estimate of the linear-Gaussian model is understandably problematic. It is known that the zeroth and first order sufficient statistics, given the hyperparameters, completely characterize the extracted i-vectors. In this study we propose to use a minimax strategy to estimate the sufficient statistics in order to increase the robustness of the extracted i-vectors. We show by experiments that the proposed minimax technique can improve over the baseline system from 9.89% to 7.99% on the NIST SRE 2010 8conv-10sec task.

Index Terms: speaker verification, minimax parameter estimate, i-vector, PLDA

1. Introduction

The recent advances in *speaker verification* methods [1], especially in the domain of *intersession variability compensation* [2, 3], have seen a breakthrough in improvements in the error rates. The chief of these techniques is the unsupervised dimenionality reduction method, known as the total variability modeling [4]. Technique is also known as an i-vector extraction. It forms a low-dimensional, fixed length, representation of the sequence of feature vectors. The i-vector can be understood to be an output of the dimensionality reduction method from mean concatenated *supervector* space to much lower dimensional, such as to 200-600 dimensions. The i-vector modeling approach appears to perform well on longer test segments (around one minute), where nuisance effects are typically in session variability, e.g. channel mismatch, additive noise, and so on.

Supervised, two-class in the case of speaker verification, classification is built on top of the extracted i-vectors, i.e. i-vectors are considered as features for the classifier. Classifiers, such as *probabilistic linear discriminant analysis* (PLDA) [5], are integrated with the intersession variabilty compensation. In a very short duration test segments, it was shown in [6] that i-vectors with PLDA scoring did not significantly outperform the *joint factor analysis* (JFA) [2]. However, cosine scoring of i-vectors [4, 7] with intersession variabilty compensation using *linear discriminant analysis* (LDA), *within-class covariance normalization* (WCCN) [4] and *nuisance attribute projection* (NAP) [3] did not outperform PLDA scoring on the short duration test segments [6]. That being the case, we will now use solely PLDA as the back-end classifier.

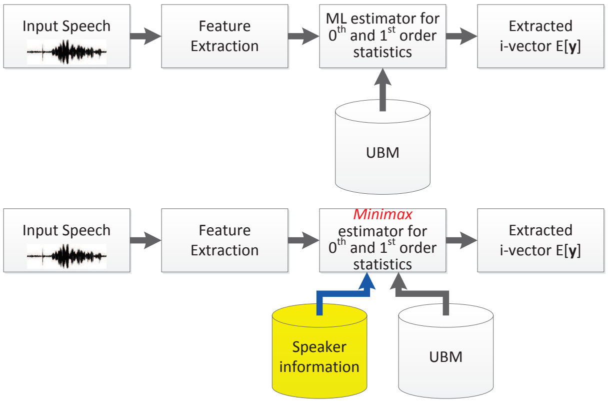
Input Speech
Feature Extraction
ML estimator for 0th and 1st order statistics
Extracted i-vector E[**y**]
UBM
Input Speech
Feature Extraction
Minimax estimator for 0th and 1st order statistics
Extracted i-vector E[**y**]
Speaker information
UBM

Figure 1: The classical (upper figure) and the proposed approach (lower figure) to the i-vector extraction.

Even with the PLDA as a classifier and session compensation technique, the performance of the state-of-the-art system degrades from 3.13% to 22.66% *equal error rate* (EER) when the test segment length was reduced from longer than one minute test segment to 2 seconds [6]. Keeping in mind, that in access control type use cases average utterace length is only 3.2 seconds [8]. It is noted that the principal challenge in achieving a low error rate is that the intra-speaker variability in the estimated parameters increases considerably as a result of variability in the lexicon and the test segment duration [8].

Utterance length has a direct relation to the increased uncertainty of the i-vector point estimate. Only the zeroth order Baum-Welch statistics, i.e. the probabilistic counts, define the covariance matrix of the posterior distribution given the utterance. When the utterance is short, the resulting posterior distribution will be wide and on the other hand when utterance is long, posterior becomes sharper [9]. One way to overcome the problem of sparsity in zeroth order statistics is to artificially add, by sampling, more probabilistic counts to the zeroth order statistics [9]. Another approach to reduce the uncertainty in i-vector estimate is extend the generative PLDA model by the Cholesky decompostion of the i-vector posterior covariance matrix [10].

In this work, we will approach the problem of the uncertainty in the sufficient statistics estimation by borrowing a technique from robust statistics [11]. We assume that nuisance attribute is generated to estimated zeroth and first order Baum-Welch statistics by just the fact that the utterance is short. A *minimax optimal* parameter estimator then finds the estimate

[*] This work was done while Ville Hautamäki was visiting Georgia Tech.

that minimizes the maximal risk among all possible estimators, i.e. it tries to counteract the maximal effect of any nuisance. A minimax optimal classifier is such a rule that will minimize the maximum classification error probability [12]. In Fig. 1 we show that in classical i-vector estimation approach statistics estimated in a maximum likelihood way, but in the minimax estimator we use the fact that the estimated i-vector will be matched against target speakers, not necessarily all, in the corpus. Originally, the minimax classification approach was used in the digit recognition task, where it utilized the fact that there are only limited number of target digits that need to be recognized. The proposed method was able to improve the baseline from 9.89% to 7.99% equal error rate.

2. Total variability modeling

In the following, we will shortly review the i-vector extraction as far as it pertains to our proposed method. The i-vector extraction is grounded on the *universal background model* (UBM), which is a C component Gaussian mixture model. It is parametrized by $\{w_c, \mathbf{m}_c, \mathbf{\Sigma}_c\}, c = 1, \dots, C$, where we have mixture weight, mean vector and covariance matrix, respectively. In this work we restrict covariance matrix to be diagonal. The well known i-vector model, a factor analysis model, is defined for the UBM component c as [4]:

$$\mathbf{s}_c = \mathbf{m}_c + \mathbf{V}_c\mathbf{y} + \epsilon_c, \tag{1}$$

where $\mathbf{V}_c$ is the sub-matrix of the total variability matrix, $\mathbf{y}$ is the latent vector, called an i-vector, ϵ_c is the residual term and $\mathbf{s}_c$ is the c'th sub-vector of the utterance dependent supervector. The ϵ_c is distributed as $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_c)$, where $\mathbf{\Sigma}_c$ is a diagonal matrix.

We then denote by $X = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ the set of feature vectors computed from the utterance. Given the set of feature vectors, the UBM and mapping from each feature vector to UBM component that generated it, we can compute the log-likelihood of the utterance by [13]:

$$\begin{aligned} &\sum_c (N_c \ln \frac{1}{(2\pi)^{F/2}|\mathbf{\Sigma}_c|^{1/2}} \\ - \quad &\frac{1}{2}\sum_t (\mathbf{x}_t - \mathbf{V}_c\mathbf{y} - \mathbf{m}_c)^{\mathsf{T}} \mathbf{\Sigma}_c^{-1} (\mathbf{x}_t - \mathbf{V}_c\mathbf{y} - \mathbf{m}_c)), \end{aligned} \tag{2}$$

where N_c is the number of feature vectors mapped to c-th component. The inner sum, indexed by t, goes over all feature vectors that are mapped to component c.

Such a map, from one frame to Gaussian component that generated it, is not directly observable. What can be done is to attach a latent binary, 1-hot, vector $\mathbf{i}_t$ of dimensionality C for each observable frame $\mathbf{x}_t$. Position of 1 in $\mathbf{i}_t$ signifies the component that generated it. We can concatenate all $\mathbf{i}$'s into a matrix $\mathbf{H}$. In this work, we approximate $\mathbf{i}_i$ by $\gamma_t(c)$ which is a posterior probability that $\mathbf{x}_t$ was generated by UBM component c:

$$\gamma_t(c) = \frac{w_c\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \mathbf{\Sigma}_c)}{\sum_{i=1}^{C} w_i\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)}. \tag{3}$$

Such an approximation is well motivated from the variational Bayes point of view [14].

Using the $\gamma_t(c)$ we can then estimate first and second order sufficient statistics, these are commonly known as Baum-Welch statistics:

$$N_c = \sum_{t=1}^{T} \gamma_t(c) \tag{4}$$

and

$$\mathbf{F}_c = \sum_{t=1}^{T} \gamma_t(c)\mathbf{x}_t, \tag{5}$$

where summation now goes over the whole set X.

Assuming that the prior of $\mathbf{y}$ is standard Gaussian, posterior, which is also Gaussian, can be computed as follows [15]:

$$\mathsf{Cov}(\mathbf{y}, \mathbf{y}) = \left(\mathbf{I} + \sum_c N_c \mathbf{V}_c^{\mathsf{T}} \mathbf{\Sigma}_c^{-1} \mathbf{V}_c\right)^{-1} \tag{6}$$

and

$$\mathsf{E}[\mathbf{y}] = \mathsf{Cov}(\mathbf{y}, \mathbf{y}) \sum_c \mathbf{V}_c^{\mathsf{T}} \mathbf{\Sigma}_c^{-1} \hat{\mathbf{F}}_c, \tag{7}$$

where $\hat{\mathbf{F}}_c$ are the centralized first order statistics, $\hat{\mathbf{F}}_c = \mathbf{F}_c - N_c\mathbf{m}_c$. We see that exctracted i-vector $\mathsf{E}[\mathbf{y}]$ is completely characterized by the Baum-Welch sufficient statistics N_c and $\mathbf{F}_c$, computed from the test utterance, and hyper-parameters $\mathbf{V}_c$, $\mathbf{m}_c$ and $\mathbf{\Sigma}_c$. In particular, $\mathsf{Cov}(\mathbf{y}, \mathbf{y})$ is dependent only on the zeroth order statistics. Small number of frames will lead to wider $\mathsf{Cov}(\mathbf{y}, \mathbf{y})$, leading to increased uncertainty of the posterior mean.

Keeping this in mind, the way to improve on the i-vector estimate is to improve on the sufficient statistics estimates. It is precisely the approach proposed here.

3. Minimax i-vector estimation

3.1. Minimax classification rule

Recalling that Baum-Welch sufficient statistics are approximations of the true zeroth and first order statistics, we will view the estimation of the statistics from the point of view of robust statistics. When there is mismatch between train and test sequences, the classification error will be adversely affected. In the robust statistics approach, a way rigorously addressing this effect is to minimize the worst case mismatch within some classes [12]. This formulation is is called the *minimax approach* [11].

Let $p_{\boldsymbol{\lambda}}(.)$ be a parameteric probability distribution and $\boldsymbol{\lambda}_i \in \Lambda$, be vector of parameters of a i-th source, $i = 1, \dots, M$. The Λ is the set of all possible parameters. Each source corresponds to one target speaker, all of the training data has to be present at the enrollment time. Potentially, i will index all possible speakers, in the present work we approximate it by a list of target speakers in the evaluation corpus. This approximation will match the out-of-set speaker to one target speaker in the corpus. We do not see this as a deficiency as speaker detection is not performed on $\boldsymbol{\lambda}$'s but on the i-vectors extracted using the $\boldsymbol{\lambda}$ estimates. As in [12], the Λ is divided into non-overlapping subsets $\Lambda_i, i = 1, \dots, M$, where $\boldsymbol{\lambda}_i \in \Lambda_i$ for all i. In [12] Λ_i is denoted as the *mismatch neighbourhood* of $\boldsymbol{\lambda}_i$.

Let the decision rule Ω be a partitioning of the whole set of possible test sequences $\mathcal{X}$ into disjoint regions $\Omega_1, \Omega_2, \dots, \Omega_M$, such that for a set of test vectors X, speaker i is recognized if $X \in \Omega_i$. Using this machinery, we can define a worst case probability of error decision rule Ω [12]:

$$p_{\Omega}(e) = \sum_{i=1}^{M} p_i \max_{\boldsymbol{\lambda} \in \Lambda_i} \int_{\Omega_i^c} p_{\lambda}(X)dX, \tag{8}$$

Where p_i is the prior probability of observing speaker i and Ω_i^c is complement of Ω_i and $p_{\boldsymbol{\lambda}}(X)$ is a parametric probability density function, parametrized by $\boldsymbol{\lambda}$. For simplicity, we will consider the flat prior, i.e. $p_i = 1/M$.

Unfortunately, direct minimization of (8) is in general not trivial [12]. However, a asymptotic approximation exists that gives a implementable decision rule [12]:

$$\Omega_i^* = \max_{1\leq j\leq M}\left[p_j \max_{\boldsymbol{\lambda}\in\Lambda_j} p_{\boldsymbol{\lambda}(X)}\right], \tag{9}$$

Next we will describe how to implement (9) in our application.

3.2. Mismatch neighbourhood

The goal is to gain a better estimate of first order statistics $\mathbf{F}_c, c = 1,\ldots,C$. In that end, we formulate a probabilistic model of $p_{\boldsymbol{\lambda}}(.)$ that will enable us to obtain $\mathbf{F}_c$ in maximum likelihood way.

The probability of one frame given UBM component c is $p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$. We will attach $\mathbf{i}$, as in Section 2, to indicate the component that generated $\mathbf{x}$. The whole utterance likelihood is $p(X|\mathbf{H},\boldsymbol{\mu},\boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are concatented mean vectors and covariance matrices, respectively. In the present work, we will consider $\boldsymbol{\mu}$ to be the parameter of interest and $\boldsymbol{\Sigma}$ to be copied from the UBM. The $\mathbf{H}$ is the latent indicator variable matrix as described in Section 2.

We write log-likelihood of $p(X|\mathbf{H},\boldsymbol{\mu},\boldsymbol{\Sigma})$ as,

$$\log p(X|\mathbf{H},\boldsymbol{\mu}) = \sum_{t=1}^{T}\sum_{c=1}^{C} i_{tc}\log p(\mathbf{x}_t|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c). \tag{10}$$

The i_{tc}, as exlained earlier, is not observable, so we approximate it by $\gamma_t(c)$. Optimization is now performed on (10).

Mismatch neighbourhood operates on Gaussian mean vector $\boldsymbol{\mu}$, which models the *Mel-frequency cepstrum coefficients* (MFCC). It was shown in [12] that the difference between cepstral coefficients of two distinct power spectral densities from a same source (speaker), is bounded above:

$$|c_\tau - \hat{c}_\tau| \leq \frac{C\rho^\tau}{\tau}, \tag{11}$$

where c_τ is the τ-th cepstral coefficient and C is a constant. The quantity is proportional to $\tau^{-1}\rho^\tau$, for some $0 \leq \rho \leq 1$. Using (11), we will now define mismatch neighbourhood similarly as in [12]:

$$\Lambda_i = \{\boldsymbol{\mu} : |\mu_\tau - \mu_\tau^{(i)}| \leq C\tau^{-1}\rho^\tau, \tau = 1,\ldots,q\}, \tag{12}$$

where q is the number of extracted cepstral coefficients. The $C > 0$ and ρ are left as user selectable parameters. In the case of delta and double delta coefficients, we approximate the mismatch neighbourhood by (12). In our initial experiments we noticed that behaviour of the boundary was similar for the plain MFCC's, deltas and double deltas.

3.3. First order statistics estimate

In order to implement (9), we utilize the same strategy as in [12]. We first take maximum likelihood estimate of $\boldsymbol{\lambda}_i = \max_{\boldsymbol{\lambda}\in\Lambda_i} p_{\boldsymbol{\lambda}}(X)$, where $\boldsymbol{\lambda}_i$ is the set of mean vectors of the target speaker i. Target speaker that maximizes (10) is selected.

To compute the new estimate of $\mathbf{F}_c$, we rewrite the log-likelihood (10) into a more convenient form where maximization is turned into minimization of:

$$-\log p(X|\boldsymbol{\mu}) = \frac{1}{2}\sum_{c=1}^{C}\sum_{t=1}^{T}\gamma_t(c)\sum_{\tau=1}^{q}\frac{\left(x_\tau^{(t)} - \mu_\tau^{(c)}\right)^2}{\sigma_\tau^2}. \tag{13}$$

In (13), we have used the fact that UBM covariance matrices were restricted to be diagonal and latent indicator variables i_{tc} are approximated by $\gamma_t(c)$. The minimization can now be computed independently per MFCC coefficient, which is clearly just a coefficient from the normalized $\mathbf{F}_c$:

$$\mu_\tau^{(c)} = \frac{1}{N_c}\sum_{t=1}^{T}\gamma_t(c)x_\tau^{(t)}. \tag{14}$$

Then the new estimate is the constrained minimization of (13), where the constraint is an interval defined by (12) per coefficient:

$$I = [\hat{\mu}_\tau^{(c)} - C\tau^{-1}\rho^\tau, \hat{\mu}_\tau^{(c)} + C\tau^{-1}\rho^\tau], \tag{15}$$

where $\hat{\mu}_\tau^{(c)}$ is a coefficient from the mean vectors of the selected target speaker i. We notice, as is the case in [12], that unconstrained minimization of (13) with respect to $\mu_\tau^{(c)}$ is convex. Optimum is then either in the interval or in the edges. In practice, we first compute the unconstrained minimum and then keep coefficient if it falls within the interval I otherwise we fix on the closest edge.

The final new mean vector is then a MAP estimate [12], where the selected target speaker mean is the prior. We utilize relevance MAP [16] to compute the new mean:

$$\boldsymbol{\mu}_{\text{new}}^c = \frac{\alpha}{N_c}\sum_{t=1}^{T}\gamma_t(c)\mathbf{x}_t + (1-\alpha)\hat{\boldsymbol{\mu}}^{(c)}, \tag{16}$$

where $\alpha = \frac{\hat{N}_c}{\hat{N}_c + r}$. The parameter r is the relevance factor, and is set to unity in the present study. In the case of zeroth order statistics $\hat{N}_c$ we need to compute new $\hat{\gamma}_t(c)$, where instead of UBM means we use the $\hat{\boldsymbol{\mu}}^{(c)} c = 1,\ldots,C$. We copy the UBM component priors and covariances matrices.

In the case of zeroth order statistics, we do not have similar characterization as exists for the cepstral coefficint in (11). So we re-estimate $\gamma_t(c)_{\text{new}}$ by plugging in μ_{new}^c, $c = 1,\ldots,C$ and keeping the component priors and covariance matrices fixed. Then the minimax estimates of zeroth and first order statistics are:

$$N_c^{\text{new}} = \sum_{t=1}^{T}\gamma_t(c)_{\text{new}} \tag{17}$$

and

$$\mathbf{F}_c^{\text{new}} = \sum_{t=1}^{T}\gamma_t(c)_{\text{new}}\mathbf{x}_t. \tag{18}$$

The new i-vector estimates are then obtained by plugging in N_c^{new} and $\mathbf{F}_c^{\text{new}}$ $c = 1,\ldots,C$ to (6) and (7). The proposed approach in i-vector estimation is summarized in Algorithm 1.

4. Experiments

In these preliminary experiments, we used the male subset of NIST SRE 2010 corpus, namely 8 conversations in training and 10 seconds in testing (8conv-10sec) condition. In a slight abuse of the NIST SRE 2010 protocol we used the whole target speaker population ($M = 194$) in the minimax search. UBM, i-vector extractor ($\mathbf{V}_c$ matrices) and PLDA speaker subspaces were estimated from Fisher, NIST SRE04, SRE05 and SRE06, Switchboard 2 and switchboard cellp 1 & 2 corpora. All speech material used in these experiments were recorded over telephone band (8 kHz). PLDA speaker subspace size was set to

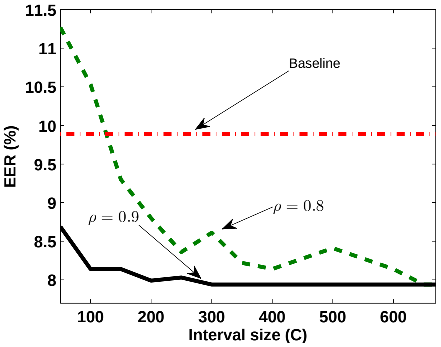

Figure 2: Varying the mismatch neighbourhood parameters of the minimax estimator.

200. From the training side all eight i-vectors were first averaged and then whitened and length normalized to unit length as in [17]. As features, we used standard 18-dimensional MFCC fearures with deltas and double deltas, leading to 54 dimensional feature space. RASTA was used. MFCC feature extraction parameters used in this study are as summarized in the Table 1. Standard energy VAD was used prior to the feature extraction.

Table 1: Summary of feature extraction parameters.

FFT size	512
Nro. Mel filters	27
Nro. cepstral coeff.	18
Frame lenght (samples)	240
Frame shift	50%

In these preliminary experiments we set the UBM size to be 512 and i-vector dimensionality to be 400. We tried also increasing the UBM size to 1024 and i-vector dimensionality to 600, which typically will improve on the error rates, in our case baseline degraded, however minimax retained more than one percent absolute improvement.

In Fig. 2 we show the stability estimator with respect to parameters C and ρ. We notice that when ρ is close to one, performance is flat with respect to interval size C. Larger the intreval, less hits to the interval edges we observe. Clearly, a significant improvement over baseline maximum likelihood estimation of first order statistics.

In table 2 we see summary of these results. Continuing the experiments from the Fig. 2, we fix the intreval size (C) to 200 and vary ρ from 0.6 to 0.9. We notice that closer to unity ρ is better sufficient statistics estimate the method is able to provide. When $\rho = 0.6$, the interval is too small and so coefficients of the mean estimate are taken from the edges of the interval. In Fig.3, we show the DET plot of the seleced experiments. We see that proposed method with $\rho = 0.9$ is clearly separated from baseline in all operating points in this plot.

Algorithm 1 Summary of the proposed approach.

```
$\mathbf{F}_c, N_c, c = 1, \ldots, C \leftarrow$ Maximum likelihood Baum-Welch statistics from $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$.
$i = \arg\max_{i=1:M} \log p(X|\boldsymbol{\mu}^{(i)})$.
$\boldsymbol{\mu}_{\text{MAP}} \leftarrow$ Compute MAP estimate using $\mathbf{F}_c, N_c, c = 1, \ldots, C$ and sufficient statistics from target speaker $i$.
Check one coeffient at the time if $\boldsymbol{\mu}_{\text{MAP}}$ is in the $\Lambda_i$. Those coefficients that are not fix to the closest edge.
Re-estimate zeroth order statistics.
Plug-in new zeroth and first order statisics estimates to (6) and (7) to obtain new i-vector estimate.
```

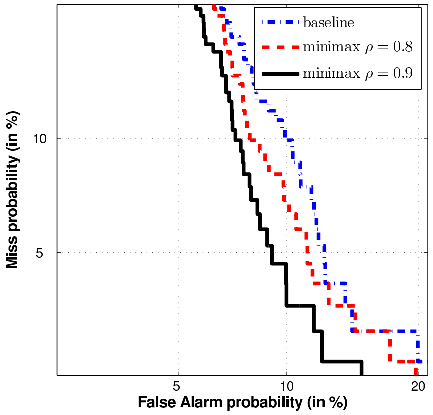

Figure 3: DET plot of systems when interval size C was kept fixed to 200. Maximum likelihood sufficient statistics estimate (baseline) result is included.

Even though absolute numbers are not among the best in the literature, we have to note that only difference between baseline and minimax is in how zeroth and first order statistics are computed. Thus, it is expected that with the improved baseline i-vector PLDA system, the proposed minimax approach can yield an improvement.

Table 2: Summary of experimental results.

Method	EER (%)	ρ	C
max. likelihood	9.89	-	-
minimax	12.21	0.6	200
minimax	8.80	0.8	200
minimax	**7.99**	0.9	200

5. Conclusions

We have proposed an i-vector estimator based on the minimax estimation of the first order statistics. In the experiments using only 10 second test samples, the proposed method was able to improve on the baseline Baum-Welch statistics estimation by a significant margin. The difference between the proposed system and the baseline system is only in how sufficient statistics were estimated.

The optimal Bayesian risk estimator with *least favorable* prior is equivalent to minimax estimator. We plan to investigate how such a Bayesian technique could be used in i-vector estimation. In addition, we plan to experiment with NIST SRE 2012 evaluation corpus as it contains variations of test segment length and realistic and digitally added noise.

6. Acknowledgements

The works of Ville Hautamäki and Padmanabhan Rajan were supported by Academy of Finland in projects 253000 and 253120, respectively.

7. References

[1] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: from features to supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12–40, January 2010.

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

[3] A. Solomonoff, W. Campbell, and I. Boardman, “Advances in channel compensation for SVM speaker recognition,” in *Proc. ICASSP 2005*, Philadelphia, Mar. 2005, pp. 629–632.

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

[5] P. Li, Y. Fu, U. Mohammed, J. Elder, and S. Prince, “Probabilistic model for inference about identity,” *PAMI*, 2011.

[6] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, “i-vector based speaker recognition on short utterances,” in *Interspeech 2011*, 2011.

[7] A. Larcher, P.-M. Bousquet, K. A. Lee, D. Matrouf, H. Li, and J.-F. Bonastre, “I-vectors in the context of phonetically-constrained short utterances for speaker verification,” in *ICASSP 2012*, 2012.

[8] A. Larcher, K. A. Lee, B. Ma, and H. Li, “The RSR2015: database for text-dependent speaker verification using multiple pass-phrases,” in *Interspeech 2012*, 2012.

[9] S. Shum, “Unsupervised methods for speaker diarization,” Master’s thesis, MIT, 2011.

[10] P. Kenny, T. Stafylakis, P. Oullet, J. Alam, and P. Dumouchel, “PLDA for speaker verification with utterances of arbitrary duratio,” in *ICASSP 2013*, Vancouver, Canada, May 2013.

[11] P. J. Huber, “A robust estimation of the probability ratio test,” *Ann. Math. Statist.*, vol. 36, no. 4, pp. 1753–1758, 1965.

[12] N. Merhav and C.-H. Lee, “A minimax classification approach with application to robust speech recognition,” *IEEE Transactions on Speech and Audio Processing*, pp. 90–100, January 1993.

[13] P. Kenny, “A small footprint i-vector extractor,” in *Speaker Odyssey 2012*, Singapore, June 2012.

[14] X. Zhao and Y. Dong, “Variational Bayesian joint factor analysis models for speaker verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 1032–1042, March 2012.

[15] 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. 345–359, May 2005.

[16] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *DSP*, vol. 10, no. 1, pp. 19–41, January 2000.

[17] D. Garcia-Romero and C. Y. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” in *Interspeech 2011*, 2011, pp. 249–252.