PLDA FOR SPEAKER VERIFICATION WITH UTTERANCES OF ARBITRARY DURATION

Patrick Kenny, Themos Stafylakis, Pierre Ouellet, Md. Jahangir Alam and Pierre Dumouchel

Centre de recherche informatique de Montréal (CRIM)

{Patrick.Kenny, Themos.Stafylakis, Pierre.Ouellet, Jahangir.Alam, Pierre.Dumouchel}@crim.ca

ABSTRACT

The duration of speech segments has traditionally been controlled in the NIST speaker recognition evaluations so that researchers working in this framework have been relieved of the responsibility of dealing with the duration variability that arises in practical applications. The fixed dimensional i-vector representation of speech utterances is ideal for working under such controlled conditions and ignoring the fact that i-vectors extracted from short utterances are less reliable than those extracted from long utterances leads to a very simple formulation of the speaker recognition problem. However a more realistic approach seems to be needed to handle duration variability properly. In this paper, we show how to quantify the uncertainty associated with the i-vector extraction process and propagate it into a PLDA classifier. We evaluated this approach using test sets derived from the NIST 2010 core and extended core conditions by randomly truncating the utterances in the female, telephone speech trials so that the durations of all enrollment and test utterances lay in the range 3–60 seconds and we found that it led to substantial improvements in accuracy. Although the likelihood ratio computation for speaker verification is more computationally expensive than in the standard i-vector/PLDA classifier, it is still quite modest as it reduces to computing the probability density functions of two full covariance Gaussians (irrespective of the number of the number of utterances used to enroll a speaker).

Index terms speaker recognition, i-vectors, PLDA

1. INTRODUCTION

The well known i-vector representation of speech segments has the convenient property that it maps segments of arbitrary duration to vectors of fixed dimension [1]. At the cost of ignoring the time dimension altogether, this representation has enabled the speaker recognition problem to be cast as an ordinary biometric pattern recognition problem like face recognition or fingerprint recognition. Numerous classifiers, all based on this simplified formulation of the problem, have been developed in recent years and shown to perform very well on the NIST speaker recognition evaluation sets [2, 3, 4, 5, 6, 7, 8, 9, 10].

However the NIST evaluation protocols simplified the speaker recognition problem in an analogous way by controlling for the durations of enrollment and test utterances in all evaluation conditions. As a result, there is little evidence available that i-vector based methods are capable of coping with the variability in utterance durations encountered in practical applications. Unlike previous NIST evaluations, this issue is addressed in the 2012 evaluation protocol where both the number of recordings available to enroll a target speaker and the durations of test utterances are allowed to vary.

In the state of the art i-vector/PLDA approach to the speaker recognition problem, an i-vector extracted from a very short utterance is treated as being just as reliable as an i-vector extracted from a long utterance. In this paper we propose to deal with the problem of duration variability by remedying this clearly unsatisfactory assumption.

Recall that the i-vector associated with an utterance is usually understood to be a point estimate of the hidden variables in an eigenvoice probability model [1, 11, 12]. We will adopt a slightly different perspective and view the i-vector as a random vector instead.

The i-vector point estimate is calculated by evaluating the posterior expectation of the hidden variables in the model conditioned on the Baum-Welch statistics extracted from the utterance. This posterior calculation provides a posterior covariance matrix as well as a posterior expectation. The posterior covariance matrix can be interpreted as quantifying the reliability of the point estimate. (It is apparent from equation (1) in [12] that the shorter the utterance, the larger this covariance matrix will be and hence the greater the uncertainty in estimating the i-vector.) Our purpose is to show how to propagate this uncertainty into the PLDA model for speaker recognition.

This uncertainty propagation can be carried out for both the Gaussian and heavy-tailed versions of PLDA [2, 3]. It is well known that i-vectors can be approximately Gaussianized by length normalization [13] so that the performance of Gaussian PLDA (with length normalization) is similar to that of heavy-tailed PLDA (without length normalization). Thus Gaussian PLDA is preferred in practice. However it is not obvious how to apply "length normalization" to posterior covariance matrices and this issue does not arise in the heavy-tailed version of PLDA. Thus we had to look into question of which version of PLDA could better accommodate posterior covariance matrices supplied by an i-vector extractor. It turned out that Gaussian PLDA yielded the best results so we restrict our attention to Gaussian PLDA in the expository portion of the paper.

2. UNCERTAINTY PROPAGATION

Given a speaker and a collection of i-vectors $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_R$ (one for each recording of the speaker), standard Gaussian PLDA assumes that the i-vectors are distributed according to

$$\boldsymbol{i}_r \quad = \quad \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{\epsilon}_r \tag{1}$$

where $\boldsymbol{m}$ is the population mean, $\boldsymbol{y}$ has a standard normal prior and the residual $\boldsymbol{\epsilon}_r$ is normally distributed with mean $\boldsymbol{0}$ and (full) covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{C}_r$ be the posterior covariance matrix associated with $\boldsymbol{i}_r$ (given by equation (1) in [12]) and let $\boldsymbol{U}_r\boldsymbol{U}_r^*$ be its Cholesky decomposition, so that the uncertainty associated

with i-vector point estimate can be expressed in the from $\boldsymbol{U}_r\boldsymbol{x}_r$ where $\boldsymbol{x}_r$ is a hidden random vector having a standard normal distribution. In this paper we modify (1) by adding this term on the right hand side:

$$\boldsymbol{i}_r \quad = \quad \boldsymbol{m} + \boldsymbol{U}_r\boldsymbol{x}_r + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{\epsilon}_r. \tag{2}$$

Thus the idea is to use the channel factors in [2, 3] to model the observation noise in the i-vector extraction process.[1] (An alternative formulation would absorb the term $\boldsymbol{U}_r\boldsymbol{x}_r$ into the residual $\boldsymbol{\epsilon}_r$. We don't take this approach since it would complicate the estimation of $\boldsymbol{\Sigma}$ in PLDA training.)

2.1. The posterior distribution of the hidden variables

Let $\underline{\boldsymbol{i}}$ denote the column vector obtained by stacking the i-vectors $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_R$ and let $\underline{\boldsymbol{X}}$ be the column vector obtained by stacking the hidden variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R, \boldsymbol{y}$. The principal computation that needs to be done in order to implement the PLDA model is to calculate the posterior distribution of $\underline{\boldsymbol{X}}|\underline{\boldsymbol{i}}$.

For $r = 1, \ldots, R$, let $\boldsymbol{F}_r$ denote the first order statistic $\boldsymbol{i}_r - \boldsymbol{m}$ and let $\underline{\boldsymbol{F}}$ be the column vector obtained by stacking $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_R$. Set

$$\underline{\boldsymbol{V}} = \begin{pmatrix} \boldsymbol{U}_1 & & \boldsymbol{V} \\ & \ddots & \vdots \\ & \boldsymbol{U}_R & \boldsymbol{V} \end{pmatrix}, \underline{\boldsymbol{\Sigma}} = \begin{pmatrix} \boldsymbol{\Sigma} & & \\ & \ddots & \\ & & \boldsymbol{\Sigma} \end{pmatrix}$$

and let

$$\underline{\boldsymbol{K}} \quad = \quad \underline{\boldsymbol{I}} + \underline{\boldsymbol{V}}^*\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{V}}.$$

Then, as in Theorem 2 of [14] The posterior $\underline{\boldsymbol{X}}|\underline{\boldsymbol{i}}$ is Gaussian with precision matrix $\underline{\boldsymbol{K}}$ and mean

$$\underline{\boldsymbol{K}}^{-1}\underline{\boldsymbol{V}}^*\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{F}}.$$

We use the notation $\langle\cdot\rangle$ to denote expectations calculated with respect to this posterior.

Like $\underline{\boldsymbol{V}}$, the matrix $\underline{\boldsymbol{K}}$ is almost, but not quite, block diagonal. Regarded as an $(R+1) \times (R+1)$ block matrix, the non-zero blocks are given by

$$\begin{aligned} \boldsymbol{K}_{rr} &= \boldsymbol{I} + \boldsymbol{U}_r^*\boldsymbol{\Sigma}^{-1}\boldsymbol{U}_r \\ \boldsymbol{K}_{r,R+1} &= \boldsymbol{U}_r^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V} \\ \boldsymbol{K}_{R+1,R+1} &= \boldsymbol{I} + R\boldsymbol{V}^*\boldsymbol{\Sigma}^{-1}\boldsymbol{V} \end{aligned}$$

for $r = 1, \ldots, R$. An algorithm which takes advantage of this type of sparsity structure to calculate Cholesky decompositions is given in Section III-D of [14]. As explained there, this gives an efficient way of calculating the first and second order posterior moments needed to implement the model, namely $\langle\boldsymbol{x}_r\rangle$, $\langle\boldsymbol{y}\rangle$, $\langle\boldsymbol{x}_r\boldsymbol{x}_r^*\rangle$, $\langle\boldsymbol{y}\boldsymbol{x}_r^*\rangle$ and $\langle\boldsymbol{x}_r\boldsymbol{x}_r^*\rangle$. Also the log evidence $\ln P(\underline{\boldsymbol{i}})$ (which useful for debugging and evaluating log likelihood ratios for speaker verification) is given by

$$\begin{aligned} &\ln \frac{1}{(2\pi)^{N/2}|\underline{\boldsymbol{\Sigma}}|^{1/2}} - \frac{1}{2}\underline{\boldsymbol{F}}^*\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{F}} \\ &\qquad - \frac{1}{2}\ln|\underline{\boldsymbol{K}}| + \frac{1}{2}\langle\underline{\boldsymbol{X}}\rangle^*\underline{\boldsymbol{V}}^*\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{F}} \end{aligned} \tag{3}$$

where N is the dimension of $\underline{\boldsymbol{X}}$.

Alternatively, these calculations can be carried out by modifying the algorithms in [3]. Essentially all that is required is to change the definitions of the matrices $\boldsymbol{J}$ and $\boldsymbol{K}$ in equations (13) and (14) of that paper (our $\boldsymbol{K}_{rr}$ and $\boldsymbol{K}_{r,R+1}$) so as to take account of the fact that the matrix $\boldsymbol{U}$ varies from one recording to another, and take the residual covariance matrix (our $\boldsymbol{\Sigma}$) to be full rather than diagonal.

2.2. Maximum likelihood estimation

So far we have only considered the case of a single speaker. In order to estimate the model parameters we need a training set comprising multiple speakers. In what follows, for a given training speaker s, r ranges over all of the recordings of the speaker and the number of these recordings is denoted by $R(s)$. We assume that the matrices $\boldsymbol{U}_r(s)$ are given. The obvious way to estimate the mean i-vector $\boldsymbol{m}$ is by averaging over the training set. So we concentrate on estimating $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$. Setting

$$\boldsymbol{\epsilon}_r(s) \quad = \quad \boldsymbol{F}_r(s) - \boldsymbol{V}\boldsymbol{y}(s) - \boldsymbol{U}_r(s)\boldsymbol{x}_r(s),$$

the EM auxiliary function for estimating $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ is

$$-\frac{R}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_s\sum_{r=1}^{R(s)}\left\langle\boldsymbol{\epsilon}_r^*(s)\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}_r(s)\right\rangle$$

where $R = \sum_s R(s)$. $\boldsymbol{V}$ is re-estimated by setting the derivative of the auxiliary function to zero so that

$$\sum_s\sum_{r=1}^{R(s)}\langle\boldsymbol{\epsilon}_r(s)\boldsymbol{y}^*(s)\rangle \quad = \quad 0.$$

The maximum likelihood estimate for $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} \quad = \quad \frac{1}{R}\sum_s\sum_{r=1}^{R(s)}\langle\boldsymbol{\epsilon}_r(s)\boldsymbol{\epsilon}_r^*(s)\rangle$$

The posterior moments $\langle\boldsymbol{\epsilon}_r(s)\boldsymbol{y}^*(s)\rangle$ and $\langle\boldsymbol{\epsilon}_r(s)\boldsymbol{\epsilon}_r^*(s)\rangle$ are evaluated using the methods in Section 2.1.

2.3. Likelihood ratio calculation for speaker verification

Given enrollment i-vectors $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_R$ for a target speaker and a test i-vector $\boldsymbol{i}$ the likelihood ratio for speaker verification is traditionally formulated as

$$\frac{P(\boldsymbol{i}_1, \ldots, \boldsymbol{i}_R, \boldsymbol{i})}{P(\boldsymbol{i}_1, \ldots, \boldsymbol{i}_R)P(\boldsymbol{i})}$$

where each term in this expression is evaluated as in (3) [2]. In the context of the NIST 2012 speaker recognition evaluation where R may be very large (up to 64), a more efficient procedure is to re-write this ratio as

$$\frac{P(\boldsymbol{i}|\boldsymbol{i}_1, \ldots, \boldsymbol{i}_R)}{P(\boldsymbol{i})} \tag{4}$$

and observe that the predictive distribution $P(\cdot|\boldsymbol{i}_1, \ldots, \boldsymbol{i}_R)$ can be viewed as another PLDA model whose parameters depend on the given speaker. Indeed, for a collection of i-vectors $\underline{\boldsymbol{i}}$ other than $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_R$,

$$P(\underline{\boldsymbol{i}}|\boldsymbol{i}_1, \ldots, \boldsymbol{i}_R) \quad = \quad \int P(\underline{\boldsymbol{i}}|\boldsymbol{y})P(\boldsymbol{y}|\boldsymbol{i}_1, \ldots, \boldsymbol{i}_R)\mathrm{d}\boldsymbol{y}$$

[1] Thanks to Niko Brümmer for making his MATLAB implementation of [3] available to us. We used this to implement uncertainty propagation in Gaussian PLDA.

where the posterior $P(\boldsymbol{y}|\boldsymbol{i}_1,\ldots,\boldsymbol{i}_R)$ is a non-standard normal distribution whose mean mean $\langle \boldsymbol{y} \rangle$ and precision matrix $\boldsymbol{P}$ are given by the methods outlined in Section 2.1; explicitly,

$$\begin{aligned} \boldsymbol{P} &= \boldsymbol{K}_{R+1,R+1} - \sum_{r=1}^{R} \boldsymbol{K}_{r,R+1}^{*} \boldsymbol{K}_{rr}^{-1} \boldsymbol{K}_{r,R+1} \\ \langle \boldsymbol{y} \rangle &= \boldsymbol{P}^{-1} \sum_{r=1}^{R} \left(\boldsymbol{V} - \boldsymbol{U}_r \boldsymbol{K}_{rr}^{-1} \boldsymbol{K}_{r,R+1} \right)^{*} \boldsymbol{\Sigma}^{-1} \boldsymbol{F}_r. \end{aligned}$$

The only reason why standard normal priors were required in the specification of the model (2) is that there is no gain in generality in permitting non-standard normal priors. In the case at hand, the speaker-dependent PLDA model can be brought to standard form by chosing speaker dependent parameters $\boldsymbol{m}'$ and $\boldsymbol{V}'$ in such a way that if $\boldsymbol{y}$ has a standard normal distribution then the first and second order moments of the expression $\boldsymbol{m}' + \boldsymbol{V}'\boldsymbol{y}$ are consistent with those of the posterior distribution $P(\boldsymbol{y}|\boldsymbol{i}_1,\ldots,\boldsymbol{i}_R)$. This can be achieved by setting

$$\begin{aligned} \boldsymbol{m}' &= \boldsymbol{m} + \boldsymbol{V} \langle \boldsymbol{y} \rangle \\ \boldsymbol{V}' &= \boldsymbol{V}\boldsymbol{T}^{-1} \end{aligned}$$

where $\boldsymbol{T}$ is the upper triangular matrix such that $\boldsymbol{T}^{*}\boldsymbol{T} = \boldsymbol{P}$ (i.e. Cholesky decomposition).

Returning to the likelihood ratio (4), it follows that both the numerator and the denominator can be evaluated using the expression for the evidence given in (3), once with the speaker-dependent PLDA model and once with the speaker-independent PLDA model.

More straightforwardly, it can be observed that the marginal distribution of a single i-vector under a PLDA model with uncertainty propagation is just a Gaussian so the likelihood ratio can be expressed as the ratio of two Gaussian pdfs. Also, some savings in computation can be obtained by noting that, although the numerator in (4) needs to be evaluated once per trial, the denominator only needs to be evaluated once per test segment. The reader may find it helpful to do the likelihood ratio calculation in detail and see how, in the case of a verification trial involving a short enrollment utterance and a short test utterance, uncertainty propagation can be expected to produce a likelihood ratio close to 1.

3. EXPERIMENTS

3.1. Evaluation data

We devised an evaluation set by randomly truncating the utterances in the female, telephone speech portion of the NIST 2010 core condition (det 5) so that the durations (after voice activity detection) of all enrollment and test utterances lay in the range 3—60 seconds. As performance metrics we used the equal error rate (EER) and the 2008 and 2010 NIST minimum detection costs (DCF_2008 and DCF_2010). Since the 2010 core condition does not contain sufficiently many trials to estimate DCF_2010 reliably, we also report results on the extended core condition (truncating enrollment and test utterances in the same way as for the core condition).

3.2. i-vector/PLDA training

As acoustic features, we used Gaussianized MFCC's (including first and second derivatives). We trained a full covariance, gender-independent UBM with 2048 Gaussians using Mixer data drawn from the 2004 and 2005 NIST speaker recognition evaluation corpora.

We trained a 600 dimensional, gender-independent i-vector extractor using the LDC releases of the Switchboard corpora, the Fisher English corpus and telephone speech data made available by NIST in 2004 and 2005; in addition, we used microphone data made available in 2004–2006 and the the interview development data made available prior to the 2008 speaker recognition evaluation. Because of the need to calculate posterior covariance matrices exactly, we used the method in [11] to train the i-vector extractor rather than the more recent method [12].

Except for the Fisher corpus, we used the same data for LDA and PLDA training. We used a 600×200 LDA projection matrix $\boldsymbol{L}$. This acts on pair $(\boldsymbol{i}, \boldsymbol{C})$ consisting of an i-vector $\boldsymbol{i}$ and a posterior covariance matrix $\boldsymbol{C}$ by

$$\begin{aligned} \boldsymbol{i} &\rightarrow \boldsymbol{L}\boldsymbol{i} \\ \boldsymbol{C} &\rightarrow \boldsymbol{L}\boldsymbol{C}\boldsymbol{L}^{*}. \end{aligned}$$

If the issue of length normalization is ignored, the corresponding matrix $\boldsymbol{U}$ in (2) is obtained by Cholesky decomposition of $\boldsymbol{L}\boldsymbol{C}\boldsymbol{L}^{*}$. We will return later to the question of how to "length normalize" covariance matrices.

We took the matrix $\boldsymbol{V}$ in (2) to be of full rank (that is, of dimension 200×200). We used uncertainty propagation in PLDA training as well as at run time (that is, enrollment and verification).

3.3. Results

We began by evaluating standard Gaussian PLDA with length normalization on the core condition with truncated utterances; results are given in Table 1.

Table 1. *Benchmark result on the NIST 2010 core condition with randomized durations (female det 5) obtained with length normalization but no uncertainty propagation.*

EER	DCF_2008	DCF_2010
6.8%	0.30	0.69

3.3.1. Scaling the Baum-Welch statistics

Our first priority in experimenting with uncertainty propagation was to look into the question of how best to quantify the uncertainty associated with the i-vector extraction process.

It is well known that successive acoustic observation vectors tend to be highly correlated. This is not generally considered to be an issue for conventional maximum likelihood training of UBMs but it may be problematic for any type of *maximum a posteriori* (MAP) estimation [15], such as the eigenvoice MAP calculation which is used to calculate point estimates of i-vectors as well as the corresponding posterior covariance matrices.

A crude way of investigating this issue is to experiment with scaling the zero and first order Baum-Welch statistics before presenting them to the i-vector extractor. A glance at a spectrogram suggests that a scale factor in the range 0.2–1.0 would be reasonable. It is clear from equation (1) in [12] that this sort of scaling would have a substantial effect on the posterior covariance matrices produced by the i-vector extractor (the smaller the scale factor the larger the posterior covariance).

To be consistent, scaling needs to be performed in training the i-vector extractor as well as at run time. Thus we built three i-vector extractors using scale factors 1/5, 1/3 and 1 for our first experiments with uncertainty propagation and we trained a Gaussian PLDA model (2) without length normalization in each case. Results on the core condition with truncated utterances are summarized in Table 2.

Table 2. *Core condition. Uncertainty propagation without length normalization. Various scalings of the Baum-Welch statistics.*

scale factor	EER	DCF_2008	DCF_2010
1	6.3%	0.32	0.65
1/3	**6.3%**	**0.31**	**0.55**
1/5	6.9%	0.32	0.57

Rather surprisingly, it turns out that the effect of scaling the Baum-Welch statistics is minor. Since using a scale factor of 1/3 gives a slight edge we used that in all subsequent experiments.

Interestingly the results Table 2 are slightly better than those in Table 1: uncertainty propagation without length normalization seems to be more effective than length normalization without uncertainty propagation.

3.3.2. Length normalization

In applying length normalization with uncertainty propagation we need to figure out what to do with the $\boldsymbol{U}$ matrices furnished by the i-vector extractor. A simple expedient when length normalizing an i-vector $\boldsymbol{i}$ is to multiply the corresponding matrix $\boldsymbol{U}$ by $1/\|\boldsymbol{i}\|$ (which is equivalent to multiplying the posterior covariance matrix by $1/\|\boldsymbol{i}\|^2$). As indicated by the results in Table 3 (labeled "scalar"), this leads to a small improvement in EER relative to the results in Table 2 (labeled "1/3").

Table 3. *Core condition. Uncertainty propagation with different length normalization methods.*

	EER	DCF_2008	DCF_2010
scalar	6.1%	0.31	0.55
unscented 1	5.9%	0.29	0.58
unscented 2	5.5%	0.27	0.60

Table 4. *Extended core condition. Uncertainty propagation with different length normalization methods.*

	EER	DCF_2008	DCF_2010	min_Cllr
scalar	6.8%	0.32	0.72	0.23
unscented 2	**5.9%**	**0.28**	**0.72**	**0.20**

A more standard way of dealing with this type of problem is to use some type of unscented transform. Suppose we are given a probability distribution (to wit, an i-vector posterior distribution) and a non-linear transformation (length normalization). To estimate the moments of the transformed distribution, we can draw a sample from the original distribution, transform each point in the sample and then calculate the moments of the transformed sample. This leads to the results labeled "unscented 1" in Table 3. For the "unscented 2" variant, we normalized the i-vector in the usual way and scaled the transformed covariance so that its determinant is the same as that of the original.

Since there is no clear winner in Table 3, we did a larger experiment using the extended core condition (with utterances truncated in the same way as for the core condition). Table 4 indicates that "unscented 2" is the best approach.

3.3.3. Gaussian vs. heavy-tailed PLDA with uncertainty propagation

Comparing the results in Tables 2 and 3 shows that length normalization is effective with uncertainty propagation but the improvement in performance is not as big as we expected. This raises the question of whether it might be better to avoid it and try using uncertainty propagation in heavy-tailed PLDA instead. As Table 5 indicates this approach did not prove to be successful.

Table 5. *Extended core condition. Gaussian vs. heavy-tailed PLDA with uncertainty propagation*

	EER	DCF_2008	DCF_2010
Heavy-tailed	7.5%	0.36	0.76
Gaussian	**5.9%**	**0.28**	**0.72**

4. CONCLUSION

The state of the art i-vector/PLDA approach to speaker recognition was developed by taking a hierarchical generative model of speech and implementing it in a mathematically incorrect way. The standard implementation is particularly defective in its handling of utterances of unrestricted duration since it treats all point estimates of i-vectors as being equally reliable. We have shown how the method of uncertainty propagation can remedy this defect.

The hierarchical generative model referred to here is a reformulation of JFA [16]. The model generates speech data for multiple recordings of a given speaker by first drawing a vector of speaker factors $\boldsymbol{y}$ from the standard normal prior. Next an i-vector for each recording is generated according to (1) and it is lifted to a GMM supervector. Finally speech data for each recording is generated sampling from the corresponding GMM.

Note that a correct implementation of this generative model would *not* treat the GMM supervectors for the recordings as being statistically independent, so that the i-vector extractor and the PLDA model (1) would need to be tightly integrated (they cannot be decoupled as in the standard i-vector/PLDA implementation). Moreover, a correct implementation would use posterior distributions of the i-vectors rather than point estimates [17, 14]. In this paper we have sought to fix the latter problem (but not the former) so as to retain the principal advantage of decoupling, namely that it enables the application of the length normalization trick. (A heavy tailed version of JFA might prove to be equally effective but this would probably be too unwieldy to experiment with.)

On the other hand, we have had to compromise on a secondary advantage derived from decoupling, namely the extremely fast evidence calculations which the standard i-vector/PLDA implementation supports. However the computational cost of the evidence calculation with uncertainty propagation is still quite modest as it consists of evaluating the probability density function of a full covariance Gaussian (typically of dimension 100–200).

5. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[2] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey 2010: The speaker and Language Recognition Workshop*, Brno, Czech Rebublic, June 2010.

[3] N. Brummer, "EM for Probabilistic LDA," Feb. 2010. [Online]. Available: https://sites.google.com/site/nikobrummer

[4] M. Senoussaoui, P. Kenny, N. Brummer, and P. Dumouchel, "Mixture of PLDA models in i-vector space for gender-independent speaker recognition," in *Proc. Interspeech 2011*, Florence, Italy, Aug. 2011.

[5] P. Matejka, O. Glembek, F. Castaldo, J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proceedings ICASSP*, 2011.

[6] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Proceedings ICASSP*, 2011, pp. 4832–4835.

[7] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," in *Proc. Odyssey 2010: The speaker and Language Recognition Workshop*, Brno, Czech Rebublic, June 2010.

[8] S. Cumani, N. Brummer, L. Burget, and P. Laface, "Fast discriminative speaker verification in the i-vector space," in *Proceedings ICASSP*, 2011, pp. 4852–4855.

[9] J. Villalba and N. Brummer, "Towards fully Bayesian speaker recognition: Integrating out the between speaker covariance," in *Proc. Interspeech 2011*, Florence, Italy, Aug. 2011.

[10] T. Stafylakis, P. Kenny, M. M. Senoussaoui, and P. Dumouchel, "Preliminary investigation of Boltzmann machine classifiers for speaker recognition," in *Proceedings Odyssey Speaker and Language Recognition Workshop*, 2012.

[11] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, pp. 345–359, May 2005.

[12] P. Kenny, "A small footprint i-vector extractor," in *Proc. Odyssey 2012*, Singapore, June 2012. [Online]. Available: http://www.crim.ca/perso/patrick.kenny

[13] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech 2011*, Florence, Italy, Aug. 2011.

[14] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms, Tech. Report CRIM-06/08-13," 2005. [Online]. Available: http://www.crim.ca/perso/patrick.kenny

[15] T. Stafylakis, P. Kenny, V. Gupta, and P. Dumouchel, "Compensation for inter-frame correlations in speaker diarization and recognition," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[16] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 16, no. 5, pp. 980–988, July 2008. [Online]. Available: http://www.crim.ca/perso/patrick.kenny

[17] X. Zhao and Y. Dong, "Variational Bayesian Joint Factor Analysis Models for Speaker Verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 1032–1042, Mar. 2012.