Non-linear PLDA for i-Vector Speaker Verification

Sergey Novoselov[1], *Timur Pekhovsky*[1,2], *Oleg Kudashev*[1], *Valentin Mendelev*[1,2], *Alexey Prudnikov*[1]

[1] Speech Technology Center Ltd., St. Petersburg, Russia
[2] ITMO University, St. Petersburg, Russia

{novoselov,tim,kudashev,mendelev,prudnikov}@speechpro.com

Abstract

Two approaches are presented for non-linear PLDA to be used in speaker verification. In NIST 2010 speaker recognition evaluation (SRE) tests under DET-5 conditions, the two methods and particularly their combination provided significant improvements in equal error rates and minDCF values over a standard PLDA scheme. The proposed schemes were also applied within a speaker verification system that employs DNN-based sufficient statistics calculation resulting in a 45 % reduction in minDCF relative to a conventional GMM based system

Index Terms: i-vector, PLDA, RBM, autoencoder, DDML

1. Introduction

Successful application of deep neural networks (DNN) [1], [2] in automatic speech recognition has provided a strong impetus to attempts seeking for possible gains from applying DNN to speaker recognition.

For example, DNN posteriors have been used by Y. Lei et al. [3], P. Kenny et al. [4] to derive sufficient statistics for alternative i-vectors calculation allowing to discriminate speakers at triphone level. According to Kenny et al. [4], this approach significantly outperformed a conventional UBM-TV-i-vectors scheme in speaker recognition.

This paper means to carry on with transfer of new deep learning (DL) technologies to the speaker recognition field. Specifically, it does not intend to apply DL just at a level of new i-vector extracton but rather to rise to a modelling level by replacing conventional PLDA [5] with a non-linear counterpart in an i-vector speaker recognition system.

Similar attempts have already been made in the speaker recognition community. For instance, a PLDA model was trained with Gaussian RBM (restricted Boltzmann machines) [3]; however, it was found to be less effective in terms of speaker verification than the classical PLDA. Supervised learning of the Gaussian RBM was performed in a manner similar to the classical PLDA approach i.e. with explicit usage of speaker and session labels. Such rigid PLDA-like learning schemes accommodate not more than two respective weight matrices for speaker and channel factors, and accordingly only one hidden layer.

This type of rigid learning scheme can be contrasted with a classical DNN learning scheme involving two steps, namely unsupervised pretraining of the RBM stack and discriminative fine-tuning of the DNN [1]. Although the fine tuning here is supervised and uses class labels, these are not used explicitly for solving two PLDA tasks of minimizing within-class variability and maximizing between-class variability.

This paper suggests two deep non-linear PLDA schemes which incorporate direct usage of class labels to solve the two PLDA tasks.

One should expect the potential gains from such models to have the same nature as in [7] where advantages of transition from linear neighbourhood component analysis (NCA) to non-linear NCA have been shown.

The remainder of this paper is organized as follows. Section 2 introduces one proposed non-linear PLDA scheme based on RBM and denoising autoencoder. Section 3 overviews an alternative non-linear PLDA approach based on discriminative deep metric learning. Experimental work is described in Section 4. Finally, our conclusions are presented in section 5.

2. Non-linear PLDA-1

An unconventional, non-linear PLDA scheme is presented which is referred to as *non-linear PLDA-1* below and aims at minimizing within-class variability of input i-vectors. The scheme is shown in Figure 1, **B** denoting binary hidden layer, **V** and **W** weight matrices*, i(s)* i-vector average for speaker *s* and *i(h,s)* i-vector referring to session *h* of the speaker *s*. **G** in Figure 1 reflects the fact that both parts of the input layer are real valued.

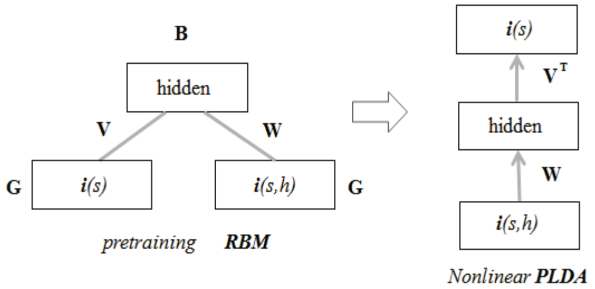

Figure 1: *Schematic diagram of non-linear PLDA-1 training.*

First, the hidden layer is pretrained as RBM [8] using standard contrastive divergence method. The learning technique is similar to that described earlier [9], except that an average of all i-vectors referring to speaker *s* is employed instead of the speaker label whereas each i-vector from the speaker *s* together with its label was used in [9].

Following pretraining, discriminative fine-tuning of the model is performed with the same data (Figure 1, right). Speaker averaged i-vectors are used as targets and individual session/speaker i-vectors as inputs for the model. The latter may be considered as a standard denoising autoencoder (DAE) [10] trained to produce the speaker i-vector while compensating for within-speaker variability (noise). MSE cost function is used at the fine-tuning stage.

The resulting autoencoder is then applied to the training set, its outputs being used as inputs to train a classical PLDA model serving as backend to provide final verification scores.

Deteriorated performance was observed in all attempts involving greater numbers of RBM-pretrained hidden layers.

3. Non-linear PLDA-2

Another non-linear PLDA scheme is based on the discriminative deep metric learning (DDML) approach introduced by Hu et al. for the facial recognition task in case of strong within-class variability [11]. The new approach named *non-linear PLDA-2* includes a deep neural net as depicted in Figure 2 and was tested as an alternative to the PDA-1.

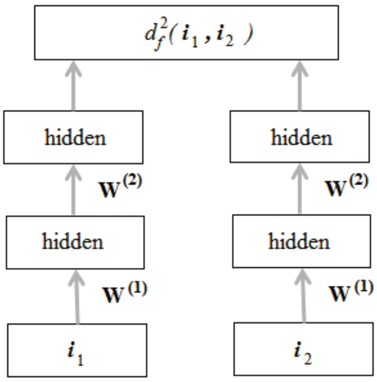

Figure 2: *Schematic diagram of non-linear PLDA-2. Here, i_1, i_2 – input i-vectors, $W^{(m)}$ – weight matrix for layer m, d^2_f – distance measure.*

In contrast to *non-linear PLDA-1,* it involves pairs of i-vectors belonging to a single speaker and pairs with an imposter to discriminate between in training a neural network model.

The trained neural network should map input i-vectors i_k and i_j into their images $f(i_k)$ and $f(i_j)$ which have to satisfy the following conditions (Figure 3):

- if input i-vectors refer to the same speaker, then $f(i_k)$ and $f(i_j)$ should lie inside a sphere of a predefined radius τ_1 (minimizing within-class variability);
- otherwise. one of $f(i_k)$ and $f(i_j)$ should lie inside an inner sphere of a spherical shell with predefined radiuses τ_1 and τ_2 where $\tau_2 > \tau_1$ and the other image should lie outside of the outer sphere.

These conditions may be expressed as follows [11]:

$$l_{kj}\left[\tau - d_f^2\left(i_k, i_j\right)\right] > R_0 \tag{1}$$

where

$$d_f^2\left(i_k, i_j\right) = \left\| f(i_k) - f(i_j) \right\|_2^2 \tag{2}$$

is Euclidean squared distance measure, $R_0 = (\tau_2 - \tau_1)/2$, $\tau = (\tau_2 + \tau_1)/2$, and pair index l_{kj} equals 1 for target pairs and -1 for imposter-involving pairs.

The objective function reflecting the required properties is $J = J_1 + J_2$ where

$$J_1 = \sum_{k,j} g\left(R_0 - l_{kj}\left[\tau - d_f^2\left(i_k, i_j\right)\right]\right) \tag{3}$$

$$g(x) = \frac{1}{\beta}\left(1 + \exp[\beta x]\right) \tag{4}$$

$$J_2 = \frac{\lambda}{2}\sum_{m=1}^{M}\left(\left\|\mathbf{W}^{(m)}\right\|_F^2 + \left\|\mathbf{b}^{(m)}\right\|_2^2\right) \tag{5}$$

Above, g is generalized logistics cost function with a parameter β, λ is a regularization constant, *M* is the number of hidden layers, $\mathbf{W}^{(m)}$, $\mathbf{b}^{(m)}$ denote weights and biases for hidden layer *m*, $\|\mathbf{W}\|_F^2$ Frobenius matrix norm. The term J_2 is introduced for regularization purposes.

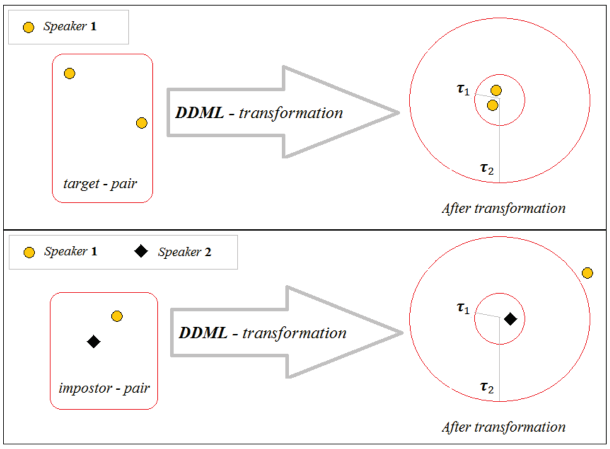

Figure 3: *Schematic diagram of the DDML approach.*

The neural net containing two hidden layers with *tanh* activation function was trained. The training set included 13000 to 16000 i-vectors. Objective function parameters R_0 and τ were equal; hence, $\tau_1 = 0$. Conventional stochastic gradient descent algorithm was employed to find weights and biases with mini-batch size of 100 pairs 50 of which were target-pairs and the rest included imposters. The pairs were chosen randomly from all available pairs of i-vectors belonging to different sessions.

More details on DDML can be found in Hu et al. [11].

4. Experiments and discussion

A JHU 2013 and DNN-based set was used in the experimentation. The former was distributed by John Hopkins University within the framework of Domain Adaptation Challenge in summer 2013 [12]. The latter was formed from different NIST corpora following the method described by Lee et al. [3].

The conditions involved were in conformance with the DET-5 extended protocol from the NIST SRE 2010 (male speakers and English language only). Equal error rate (EER) and minimum normalized detection cost function (minDCF) were used to measure the performance of the verification systems. Each system had a conventional two covariance model [13] as backend. All input vectors were centered and whitened with parameters estimated on the training set. The same parameters were used during the tests.

In Tables 1-4, RBM+PLDA and RBM+DAE+PLDA rows contain results of *non-linear PLDA-1* with pretraining only and of the full version of *non-linear PLDA-1* (pretraining and fine-tuning) correspondingly. DDML+PLDA stands for *non-linear PLDA-2* rows. RBM+DAE+DDML+PLDA denote a combination of *non-linear PLDA-1* and *non-linear PLDA-2*.

4.1. Experiments with JHU-2013 set of i-vectors

Publicly available JHU-2013 i-vector set was elected to avoid frontend dependence which could impair reproducibility of results. The i-vector dimensionality was 600, as recommended [12].

The data used for training included MIXER-set telephone speech from NIST SRE 2004-2008 corpora (13626 sessions from 1115 male speakers, English only).

Results obtained with NIST SRE 2010 are given in Table 1. The upper row shows the results for the PLDA backend on i-vectors. The second and the third row shows the results for *non-linear PLDA-1* before and after the fine-tuning stage to emphasize its contribution to the system performance.

As the fourth row shows, *non-linear PLDA-2* is inferior to *non-linear PLDA-1* but both of them outperform the reference system by a significant margin which may be termed non-linearity gap.

Table 1. *Tests involving JHU-2013 set.*

System	EER (%)	minDCF
Baseline PLDA	2.17	0.362
RBM+PLDA	2.16	0.372
RBM+DAE+PLDA	1.69	0.315
DDML+PLDA	1.97	0.326
RBM+DAE+DDML+PLDA	**1.65**	**0.308**

The fifth row refers to a hybrid system combining DDML with PLDA-1 which turned out to be the best for the JHU-2013 set. One hidden layer with 2000 nodes and a stack of two hidden layers having 2000 and 600 nodes were employed for the PLDA-1 and PLDA-2 respectively. Deteriorated performance was observed in all attempts involving greater numbers of hidden layers.

4.2. Experiments with DNN-based set of i-vectors

A DNN-based i-vectors set (Figure 4) was selected to experimentally investigate how the non-linearity gap would manifest itself with the new type of i-vectors [3].

Here, special attention was paid to relative error reduction; no attempt was made to top the state of the art result for DNN-based systems. One reason was that the Fischer speech corpus was not available which would be necessary to build a competitive DNN for sufficient statistics assessment on senones.

Similarly to Lei et al. [3], a comparison between DNN-based i-vectors and conventional i-vectors was attempted to estimate DNN-induced performance difference. To do it, systems other than DNN posteriors based pseudo-UBM were trained, namely:

- classical diagonal UBM with 2048 Gaussian components; the resulting i-vector set is referred to as *the standard set* below;
- UBM trained in supervised manner as described earlier, see Lei et al. [3], Section 5, yielding what is called *the supervised set* below.

Our ASR DNN was trained with KALDI toolkit [14] on the SwitchBoard RC2 speech corpus and had 2720 outputs (senones). Accordingly, the supervised UBM included 2720 components. The network comprised 6 hidden layers 2048 nodes each with sigmoid activation function. The hidden layers were generatively pretrained as RBMs. Then the network has been discriminatively fine-tuned with stochastic gradient descent algorithm.

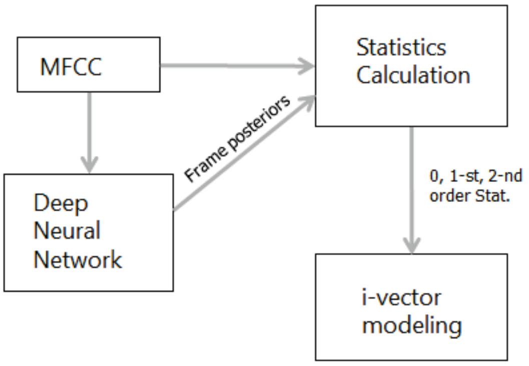

Figure 4: *Schematic diagram of DNN-based i-vector extractor.*

The training data for all components of the systems involved consisted of telephone data in English from NIST SRE 1998-2008, namely 16618 sessions corresponding to 1763 male speakers.

KALDI toolkit [14] was used to generate 60-dimensional features (20 MFCC including C0 with delta and delta-delta) for both training and testing in accordance with NIST SRE 2010 DET 5 extended conditions. The dimensionality of i-vectors was 400. We used raw i-vectors instead of LDA projections in contrast to Lei et al. [3].

Table 2. *Tests involving the standard set.*

System	EER (%)	minDCF
Baseline PLDA	2.76	0.519
RBM+PLDA	2.78	0.506
RBM+DAE+PLDA	2.31	0.490
DDML+PLDA	2.33	0.486
RBM+DAE+DDML+PLDA	**2.02**	**0.460**

For *non-linear PLDA-1*, a single hidden layer with 1300 nodes was found to be optimal. For DDML–PLDA, the first hidden layer comprised 1000 nodes and the second one 400 nodes; $R_0 = 0.3$ and $\tau = 0.3$. For the hybrid of *non-linear PLDA-1* and *non-linear PLDA-2,* $R_0 = 0.1$ and $\tau = 0.1$.

Table 3. *Tests involving the supervised set.*

System	EER (%)	minDCF
Baseline PLDA	2.66	0.428
RBM+PLDA	2.76	0.415
RBM+DAE+PLDA	2.20	0.359
DDML+PLDA	2.35	0.355
RBM+DAE+DDML+PLDA	**2.11**	**0.335**

Results are shown in Tables 2-4. The non-linearity gap may be assessed by comparing the top and bottom row in each table. For example, the DNN-based set of i-vectors (Table 4) offered 27 % and 23 % reductions in terms of EER and minDCF.

Table 4. *Tests involving the DNN-based set.*

System	EER (%)	minDCF
Baseline PLDA	1.98	0.365
RBM+PLDA	1.93	0.338
RBM+DAE+PLDA	1.61	0.301
DDML+PLDA	1.71	0.307
RBM+DAE+DDML+PLDA	**1.44**	**0.282**

Comparing top rows of Tables 2, 3 and 4 yields performance gain due to the DNN-based i-vectors extractor (DNN-gap); the difference between *the standard* and DNN-based sets of i-vectors is 28 % for EER and 29 % for minDCF.

Finally, the difference between the top row of Table 2 and the bottom row of Table 4 represents an overall advantage of the DNN-based system combining *non-linear PLDA-1* and *non-linear PLDA-2* over the standard approach. Specifically, the performance yield is as high as 48 % in terms of EER and 45 % in terms of minDCF.

5. Conclusions

Two schemes of non-linear PLDA for i-vectors were introduced. Each of the schemes alone and the two schemes combined significantly outperformed the standard baseline PLDA scheme in the NIST SRE 2010 DET 5 testing conditions. The speaker verification system that employs DNN-based sufficient statistics calculation and the combination of the proposed schemes achieved 45 % relative minDCF reduction by comparison with a conventional GMM based system.

We plan to further develop the presented approach by enabling joint discriminative training of non-linear projectors together with a discriminative PLDA [15] backend.

6. Acknowledgements

This work was in part financially supported by the Government of the Russian Federation, Grant 074-U01.

7. References

[1] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[2] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[3] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically aware deep neural network,” in *Proc. 2014 IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 1695–1699.

[4] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam. Deep Neural Networks for extracting Baum-Welch statistics for Speaker Recognition. presented at Odyssey 2014: *The Speaker and Language Recognition Workshop* [Online]. Available: http://cs.uef.fi/odyssey2014/program/pdfs/28.pdf

[5] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proc. IEEE 11th Int. Conf. Comput. Vision*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[6] T. Stafylakis, P. Kenny, M. Senoussaoui, and P. Dumouchel, “PLDA using Gaussian Restricted Boltzmann Machines with application to Speaker Recognition,” in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, Portland, OR, USA, 2012, pp. 1692–1696.

[7] R. Salakhutdinov and G. Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. presented at AISTATS 2007 [Online]. Available: http://www.cs.toronto.edu/~fritz/absps/nonlinnca.pdf

[8] H. Larochelle and Y. Bengio, “Classification using discriminative restricted Boltzmann machines”, in *Proc. 25th Int. Conf. Machine Learning*, Helsinki, Finland, 2008, pp. 536–543.

[9] S. Novoselov, T. Pekhovsky, and K. Simonchik. STC Speaker Recognition System for the NIST i-Vector Challenge. presented at Odyssey 2014: *The Speaker and Language Recognition Workshop* [Online]. Available: http://cs.uef.fi/ odyssey2014/program/pdfs/25.pdf

[10] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol. Extracting and composing robust features with denoising autoencoders. presented at 25th Int. Conf. Machine Learning, Helsinki, Finland, 2008 [Online]. Available: http:// icml2008.cs.helsinki.fi/papers/592.pdf

[11] J. Hu, J. Lu, and Y. Tan. Discriminative Deep Metric Learning for Face Verification in the Wild. presented at 2014 IEEE Conf. Comput. Vision and Pattern Recognition (CVPR) [Online]. Available: http://www.cv-foundation.org/openaccess/ content_cvpr_2014/papers/Hu_Discriminative_Deep_Metric_2014_CVPR_paper.pdf

[12] JHU 2013 speaker recognition workshop. [Online]. Available: http://www.clsp.jhu.edu/workshops/archive/ws13-summerworkshop/%20groups/spk-13/

[13] N. Brummer and E. de Villiers. The speaker partitioning problem. presented at Odyssey 2010: *The Speaker and Language Recognition Workshop* [Online]. Available: http://sites.google.com/site/nikobrummer/brummer_odyssey10draft.pdf

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in *IEEE Automatic Speech Recognition and Understanding Workshop* , December 2011.

[15] S. Cumani, N. Brummer, L. Burget, P. Laface, O. Plchot, and V. Vasilakakis, “Pairwise discriminative speaker verification in the i-vector space,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 6, pp. 1217–1227, 2013 [Online]. Available: http://porto.polito.it/2506145/1/manuscript_R2_open.pdf