I-VECTOR KULLBACK-LEIBLER DIVISIVE NORMALIZATION FOR PLDA SPEAKER VERIFICATION

Yilin Pan, Tieran Zheng, Chen Chen

School of Computer Science and Technology, Harbin Institute of Technology, Harbin

ABSTRACT

I-vector and Probabilistic Linear Discriminant Analysis (PLDA) represents the state-of-the-art in the speaker verification system. In PLDA, the i-vectors are assumed to follow Gaussian distribution. However, this assumption results in poor modeling without Gaussianization. Different from previous Gaussianization methods, in our proposed method, we make no restriction towards the original distribution of i-vectors for flexibility and universality. To optimize the Gaussian transformation function, Kullback-Leibler divergence (KLD) is introduced to measure the distance between the two distributions. By minimizing the KLD value under the development data, we can search out the optimal parameters in transformation function. The proposed method shows significant improvement on NIST SRE 2008 core set; together with length normalization (LN), a famous Gaussianization method, can further improve the verification accuracy.

Index Terms— speaker verification, PLDA, Gaussianization, divisive normalization, Kullback-Leibler divergence

1. INTRODUCTION

Speaker verification system based on i-vectors [1] and PLDA [2] represents the current state-of-the-art and has received considerable attention in related fields [3, 4]. The most popular PLDA model, named Gaussian PLDA (G-PLDA), works under the assumption that the latent variables and the i-vectors both follow Gaussian distribution. However, it has been shown that this assumption does not hold in the presence of channel disturbance [5]. To deal with the non-Gaussian behavior of i-vectors, a heavy-tail Probabilistic Linear Discriminate Analysis (HT-PLDA) model [5] was proposed that adopts a heavy-tail distribution to take the place of the Gaussian assumption in the model. The HT-PLDA model shows superior performance for speaker recognition from a telephone channel, but it is ineffective for microphone channels because the channel effects are extreme and not follow heavy distribution. Moreover, the algorithm is computationally expensive both in training and in testing.

This research is partly supported by the National Natural Science Foundation of China under Grant Nos. 61471145 and 91120303.

To take advantage of the simplicity and computational efficiency of G-PLDA, several techniques [6–9] was proposed to fulfill the Gaussian assumption. LN [6] assumed that i-vectors' original distribution belongs to the family of Elliptically Symmetric Densities (ESD). Under this assumption, i-vectors were projected onto a spherical surface to meet the Gaussian assumption. Furthermore, iterative LN [8] was proposed to modify the i-vectors distribution, making it to approach Gaussian distribution further. The work described in [9] assumed that the i-vectors distribution follows a sinh-arcsinh distribution [10] and could be transformed by a sequence of affine and nonlinear transformations. Compared with HT-PLDA, above Gaussianization methods can achieve equivalent or preferable performance under the G-PLDA framework.

However, the above gaussianization methods are limited with regard to real-world applications due to their specific assumptions about the i-vectors' original distribution. To improve the applicability of our method, this paper presents gaussianization method named Kullback-Leibler Divisive Normalization (KL-DN) to make i-vectors satisfy the Gaussian assumption in G-PLDA. The contributions of our proposed method can be summarized as follows: (i) For flexibility and universality, KL-DN makes no assumption about the original distribution of the i-vectors. (ii) To optimize the transformation function in training phase, the KLD [11] between the i-vectors' distribution and the standard Gaussian distribution is minimized by transforming i-vectors. (iii) The results show that a negative correlation exists between the KLD and the verification accuracy. The Results also show that KL-DN achieves a superior performance compared with LN and combined with LN can further improve the accuracy.

The remainder of this paper is organized as follows. Section 2 provides a description of the related works. The proposed i-vector Gaussianization method is described in Section 3. Section 4 presents to experimental setup and the behavior of proposed method on the male portions of the core sets from NIST SRE 2008. Section 5 draws conclusions.

2. RELATED WORKS

This section we provide a brief overview of state-of-the-art speaker verification systems and a transforming method, re-

spectively.

2.1. Speaker verification system

The i-vector extraction approach was proposed in [1]. Given a speaker utterance s, the speaker-space and channel-space-dependent GMM-supervector $\boldsymbol{\theta}$ is written as follows:

$$\boldsymbol{\theta} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{x}, \tag{1}$$

where the supervector $\boldsymbol{m}$ comes from the Universal Background Model (UBM). $\boldsymbol{T}$ is a low-rank total variability (TV) matrix. The $\boldsymbol{x}$ is a D-dimension speaker identity vector called i-vector. In the Eq.(1), i-vectors are assumed to follow standard Gaussian distribution.

For a speaker i, the collection of corresponding i-vectors concerning utterance $\{j = 1, 2 \cdots, S\}$ is denoted as $\boldsymbol{x}_{ij}$. The G-PLDA model then assumes that each i-vector can be decomposed $\boldsymbol{x}_{ij}$ as follows:

$$\boldsymbol{x}_{ij} = \boldsymbol{\mu} + \boldsymbol{F}\boldsymbol{h}_i + \boldsymbol{G}\boldsymbol{w}_{ij} + \boldsymbol{\epsilon}_{ij}. \tag{2}$$

This model can be decomposed into two parts: (i) the signal component $\boldsymbol{\mu} + \boldsymbol{F}\boldsymbol{h}_i$ which describes between-speaker variability and (ii) the channel component $\boldsymbol{G}\boldsymbol{w}_{ij} + \boldsymbol{\epsilon}_{ij}$, which describes within-speaker variability. The latent identify vector $\boldsymbol{h}_i$ and the latent vector $\boldsymbol{w}_{ij}$ are assumed to statistical independence and both follow standard Gaussian distribution.

2.2. I-vector regularization

LN is often applied prior to G-PLDA to address the non-Gaussian behavior of i-vectors. Before LN, the i-vectors should be standardization [6] to transform i-vectors from the ESD family into the Spherically Symmetric Density (SSD) family . After LN, all the i-vectors are scaled to the unit length and lie on the maximum density shell of a standard Gaussian, making the i-vectors to be closer to a standard Gaussian distribution [12].

3. I-VECTOR GAUSSIANIZATION BASED ON KL-DN

From the paper [5] we can draw a conclusion that the i-vectors follow non-Gaussian distribution. In our proposed method, to meet the Gaussian assumption in G-PLDA, we propose a Gaussianization method to reduce the KLD between the i-vectors and the standard Gaussian distribution by transforming the i-vectors non-linearly. For flexibility and universality, we make no assumption about the i-vectors' original distribution.

In this section, we present a formal mathematical description of our proposed method. First, the i-vector transformation function is presented; then, the optimization function is presented which takes the role of searching out the best parameters in the transformation function.

3.1. Transformation function

Divisive normalization (DN) proposed in [13] is used in biological vision modeling. Subsequently, it has been widely used to explain human visual neurons [14–16], olfactory receptor neurons [17] and image processing field [18,19]. In our task, a non-linear transformation is required according to [6]. The DN transformation function, as a nonlinear transformation aimed at reducing the Mutual Information (MI) of vectors, has attracted extensive attention in the research fields listed above both for its simplicity and its superior performance. Thus we introduce the transformation function into out method to transforming i-vectors:

$$(x_{kldn})_d = \frac{x_d}{(b + \sum_{i=1}^{D} c_i x_i^2)^{\frac{1}{2}}}, \quad for\ d = 1, \cdots, D, \tag{3}$$

where the term $\boldsymbol{x}_{kldn}$ represents an i-vector belonging to the KL-DN transformed domain. b and c_i are the related transformation parameters. The weights are all identical ($c_i = c, i = 1, \cdots, D$) after $\boldsymbol{x}$ is whitened. For simplicity, in the preprocessing stage, the i-vectors are whitened before using Eq. (3) in our method. To Gaussianization i-vectors through Eq.(3), we propose a optimization function to search out transformation parameters b and c.

3.2. Optimization function

To search out the transformation parameters b and c in Eq.(3), we minimize the distance between the i-vectors PDF $p(\boldsymbol{x})$ and standard Gaussian $N(\boldsymbol{0}, \boldsymbol{I})$. In this paper, KLD is employed as the distance measurement method. The distance measured by KLD is named as *negentropy* and defined as $J(\boldsymbol{x})$. From Eq. (5) in [20], we know that $J(\boldsymbol{x})$ can be decomposed by considering the target PDF $N(\boldsymbol{0}, \boldsymbol{I})$:

$$\begin{aligned} J(\boldsymbol{x}) &= D_{KL}(p(\boldsymbol{x})|N(\boldsymbol{0}, \boldsymbol{I})) \\ &= D_{KL}(p(\boldsymbol{x})|\Pi_d p(x_d)) + \sum_{d=1}^{D} D_{KL}(p(x_d)|N(0,1)). \end{aligned} \tag{4}$$

In Eq. (4), the first part represents MI, expressed as I, which is used to measure the statistical dependence of i-vector elements. MI equals to zero if and only if the elements in $\boldsymbol{x}$ are independent. The second part of Eq. (4) represents marginal negentropy and is denoted by $J_m(\boldsymbol{x})$, which is used to measure the distance between every single dimension in i-vectors and standard Gaussian $N(0, 1)$. Given an unknown PDF, both $J_m(\boldsymbol{x})$ and $I(\boldsymbol{x})$ are non-negative. It can be found that $J(\boldsymbol{x}_{wht})$ is constant.

In this section, we attempts to search out the parameters b and c by minimizing the $J(\boldsymbol{x}_{kldn})$, in order to introduce b and c into optimization function, we maximize the difference between the $J(\boldsymbol{x}_{wht})$ and $J(\boldsymbol{x}_{kldn})$:

$$\max \Delta J = \max(\Delta I + \Delta J_m). \tag{5}$$

First, we provide the concrete representation of ΔI between $\boldsymbol{x}_{wht}$ and $\boldsymbol{x}_{kldn}$ as follows:

$$\begin{aligned}\Delta I =& \sum_{d=1}^{D} H((x_{wht})_d) - H(x_{wht}) \\ & - [\sum_{d=1}^{D} H((x_{kldn})_d) - H(x_{kldn})] \\ =& \sum_{d=1}^{D} H((x_{wht})_d) - \sum_{d=1}^{D} H((x_{kldn})_d) \\ & + \langle \log |det(\frac{\partial \boldsymbol{x}_{kldn}}{\partial \boldsymbol{x}_{wht}})| \rangle_{\boldsymbol{x}_{wht}},\end{aligned} \tag{6}$$

where $H(\boldsymbol{x}) = -\int p(\boldsymbol{x}) \log(p(\boldsymbol{x}))$ is the differential entropy of $\boldsymbol{x}$. Note that $\sum_{d=1}^{D} H((x_{wht})_d)$ is constant with respect to the transformation parameters and can be omitted. Here, $\langle \cdot \rangle_{\boldsymbol{x}_{wht}}$ denotes computing the expected log Jacobian for $\boldsymbol{x}_{wht}$. By considering Eq.(3), the Jacobian is represented as:

$$\det(\frac{\partial \boldsymbol{x}_{kldn}}{\partial \boldsymbol{x}_{wht}}) = \frac{b}{(b + cr^2)^{(D/2+1)}}, \tag{7}$$

where $r = \|\boldsymbol{x}_{wht}\|$ is used to represent the ℓ_2-norm of the whitened i-vectors. By substituting Eq. (7) into Eq. (6), we can rewrite Eq. (6) as

$$\Delta I \equiv -\sum_{d=1}^{D} H((x_{kldn})_d) + \log b - (\frac{D}{2} + 1)\langle \log(b + cr^2) \rangle_r. \tag{8}$$

In practice, both $H(\boldsymbol{x}_{kldn})$ and r need to be estimated from development i-vectors.

It is worth mentioning that Eq.(8) is the optimization function of DN. Different from DN, which aiming at reducing the statistical dependence of i-vector elements, we also take the difference of i-vectors' marginal negentropy ΔJ_m into consideration. Next, the second part ΔJ_m in Eq. (5) is decomposed as follows:

$$\begin{aligned}\Delta J_m \equiv & -\sum_{d=1}^{D} D_{KL}(p(x_{kldn})_d | N(0,1)) \\ = & \sum_{d=1}^{D} H((x_{kldn})_d) + \sum_{d=1}^{D} \int p((x_{kldn})_d) \log(p(N(0,1)),\end{aligned} \tag{9}$$

where $J_m(\boldsymbol{x}_{wht})$ in Eq. (9) is omitted as a constant. By combining Eq.(8) and Eq.(9) according to Eq.(5), we can obtain our optimization function:

$$\begin{aligned}\arg\max_{b,c} \Delta J =& \sum_{d=1}^{D} \int p((x_{kldn})_d) \log(p(N(0,1))) \\ & + \log b - (\frac{D}{2} + 1)\langle \log(b + cr^2) \rangle_r.\end{aligned} \tag{10}$$

Through optimization function, the related parameters b and c of the transformed function are calculated through a grid search. It is worth mentioning that all the related PDFs in Eq.(10) are estimated from the development data in training phase by a non-parametric statistical method to maintain the flexibility of the proposed method. In test phase, already trained Eq.(3) is applied on target i-vector and test i-vector for Gaussianization.

4. EXPERIMENTS AND ANALYSIS

This section presents experimental validations of the effectiveness of KL-DN in speaker verification performance. The following section provides details about the experimental setup and the results of the proposed approach.

4.1. Experimental setup

Our experiments were performed on the male portion of the core sets from NIST SRE 2008 and it was refered as the evaluation data. We report the results of trials through interview-interview speech(det1) and core set(det1-det8). For a performance metric, we used the equal error rate (EER) and DCF to the minimum value of the 2008 NIST detection cost function (minDCF08).

In all the experiments, the sentences were represented by a 60-dimensional vector of Mel Frequency Cepstral Coefficients (MFCC), which was extracted using a 25 ms Hamming window with a 10 ms frame advance. In particular, 20 MFCC together with their first and second derivatives compose the MFCC feature. A full-covariance gender-dependent UBM with 2,048 mixtures was trained from NIST SRE 2003–2006 which is referred as development data. The dimension of the gender-dependent i-vector extractor is 600. before i-vectors' Gaussianization, LDA was used to project the i-vectors into 120 dimensions.

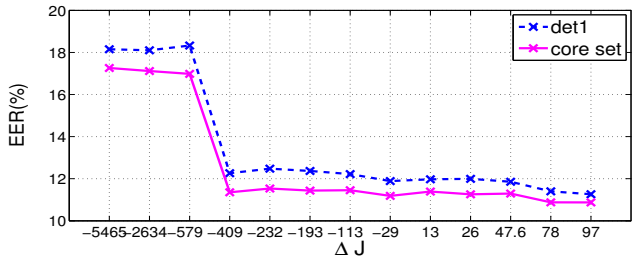
4.2. Optimization function analysis

Table 1 summarizes the value of the optimization function and the related transformation parameters of the optimization function under the KL-DN and DN methods. The value of ΔJ in DN was derived under the transformation parameters which were obtained by maximizing ΔI.

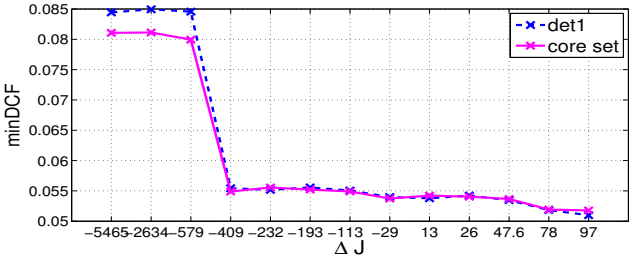
Table 1. The results of the optimization functions.

Method	trans-parameter		opti-function	
	$\log_2 b$	$\log_2 c$	ΔI	ΔJ
DN	-19	-20	-512.4	$-5,451.0$
KLDN	-12	-13	$-$	97.8

From Table 1, it can be found that the ΔJ under DN is negative, which reveals that maximizing MI can only enlarge



(a) Variation between EER(%) and ΔJ



(b) Variation between minDCF and ΔJ

Fig. 1. Variation between the verification results and ΔJ under the conditions in the male portion of the det1 and core sets (short2–short3) of NIST SRE 2008.

the distance between i-vector and standard Gaussian distributions. However, taking KLD as the optimization function can search out the parameters that narrows the distance between the original and target distributions. The results explain that controlling MI is not enough to Gaussianization i-vectors, marginal negentropy should also be taken into consideration.

Figure 1 displays the relationship between ΔJ derived on development data and the speaker verification results derived on evaluation data. We chose the transformation parameters b and c randomly and plotted the corresponding ΔJ on the horizontal coordinate axis. The verification results are represented in terms of EER and minDCF08. The two graphs prove that narrowing the distance by our proposed method can improve speaker verification accuracy.

4.3. Verification results analysis

The experimental results were compared under the G-PLDA system with i-vector Gaussianization methods, including LN and KL-DN with and without LN. The transformation parameters for KL-DN are listed in the results in Table 1 ($log_2 b = -13, log_2 c = -12$). The results are summarized in Table 2 and related DET curves are shown in Figure 2. The results, in terms of the evaluation criterion of EER and minDCF08, are summarized in Table 2 and related DET curves are shown in Figure 2.

By comparing the results of the three different approaches, we can summarize as follows:

(i) Under the det1, KL-DN achieves a relative improvement of 8.9% in terms of minDCF08 over the baseline, decreasing from 0.056 to 0.051; and a relative improvement of 6.7% in EER over the baseline, decreasing from 12.07% to 11.26%. It proves the effectiveness of our proposed method under the interview channel condition.

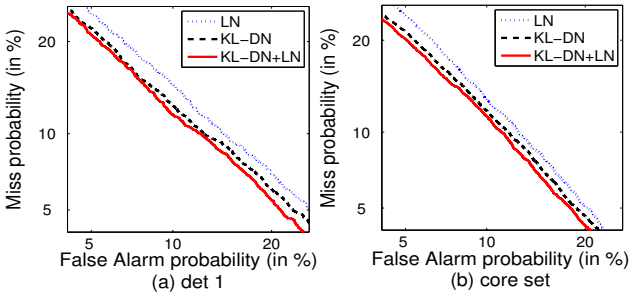


Fig. 2. The DET curves obtained with LN, KL-DN with and without LN in the male portion of the det1 and core sets (short2–short3) of NIST SRE 2008..

(ii) To prove the effectiveness of the proposed method in a variety of conditions, we tested it on the entire core set, which includes eight different conditions. The results show clearly that our proposed method achieves superior improvements in terms of both minDCF08 and EER.

(iii) Using the same transformation parameters, KL-DN together with LN exhibits further improvement. We presume that KL-DN reduces the non-Gaussian behavior in i-vectors and then decreases the mismatching of i-vectors distribution and the assumption in LN; a theoretical explanation will be provided in future work.

Table 2. Speaker verification results under the conditions in the male portion of the det 1 and core sets (short2–short3) from NIST SRE 2008.

System code	det1		core set	
	EER (%)	minDCF	EER(%)	minDCF
LN	12.07	0.056	11.40	0.055
KL-DN	11.26	**0.051**	10.87	0.052
KL-DN+LN	**10.87**	**0.051**	**10.58**	**0.051**

5. CONCLUSIONS

In this paper, we propose a Gaussianization method to transform the distribution of i-vectors. We prove that it can improve the speaker verification performance by narrowing the KLD between two distributions through promoted transformation function. Moreover, combining LN and KL-DN can further improves the verification accuracy.

6. REFERENCES

[1] N. Dehak, P. J. 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.

[2] S. J.D̃. Prince and J.H̃. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings of IEEE International Conference on Computer Vision 2007*, 2007, pp. 1–8.

[3] John H. L. Hansen and Taufiq Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[4] Amir Nemat, "Distant speaker recognition an overview," *International Journal of Humanoid Robotics*, vol. 13, no. 02, 2015.

[5] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proceedings of Odyssey 2010: Speaker and Language Recognition Workshop*, 2010.

[6] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH 2011, Florence, Italy, August*, 2011, pp. 3283–3291.

[7] Pierre Michel Bousquet, Driss Matrouf, and Jean Franois Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *INTERSPEECH 2011, Florence, Italy, August*, 2011, pp. 485–488.

[8] P. M. Bousquet, A. Larcher, D. Matrouf, J. F. Bonastre, and O. Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis," in *Proceedings of Odyssey 2012: Speaker and Language Recognition Workshop*, 2012.

[9] Sandro Cumani and Pietro Laface, "I-vector transformation and scaling for PLDA based speaker recognition," in *Proceedings of Odyssey 2016: Speaker and Language Recognition Workshop*, 2016, pp. 39–46.

[10] M. C. Jones and Arthur Pewsey, "Sinh-arcsinh distributions," *Biometrika*, vol. 96, no. 4, pp. 761–780, 2009.

[11] Thomas M. Cover and Joy A. Thomas, *Elements of information theory (2. ed.)*, Wiley, 2006.

[12] Pierre Michel Bousquet and Jean Francois Bonastre, "Constrained discriminative speaker verification specific to normalized i-vectors," in *Proceedings of Odyssey 2016: Speaker and Language Recognition Workshop*, 2016, pp. 53–59.

[13] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Visual Neuroscience*, vol. 9, no. 2, pp. 181–197, 1992.

[14] M Carandini and D. J. Heeger, "Normalization as a canonical neural computation," *Nature Reviews Neuroscience*, vol. 13, no. 1, pp. 51–62, 2011.

[15] H. H. Li, M Carrasco, and D. J. Heeger, "Deconstructing interocular suppression: Attention and divisive normalization," *Plos Computational Biology*, vol. 11, no. 10, pp. 1–26, 2015.

[16] Tatsuo K Sato, Haider Bilal, Husser Michael, and Carandini Matteo, "An excitatory basis for divisive normalization in visual cortex," *Nature Neuroscience*, vol. 19, no. 4, pp. 568–570, 2016.

[17] Shawn R Olsen, Vikas Bhandawat, and Rachel I Wilson, "Divisive normalization in olfactory population codes," *Neuron*, vol. 66, no. 2, pp. 287–99, 2010.

[18] Siwei Lyu, "Divisive normalization: Justification and effectiveness as efficient coding transform," in *Proceedings of Conference on Neural Information Processing Systems 2010*, 2010, pp. 1522–1530.

[19] Qian Xu and L. J Karam, "Change detection on sar images using divisive normalization-based image representation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 4339–4343.

[20] Jean Fran Cardoso, "Dependence, correlation and gaussianity in independent component analysis," *Journal of Machine Learning Research*, vol. 4, no. 4, pp. 1177–1203, 2003.