CONSTRUCTION OF DISCRIMINATIVE KERNELS FROM KNOWN AND UNKNOWN NON-TARGETS FOR PLDA-SVM SCORING

Wei Rao and Man-Wai Mak

Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hong Kong SAR, China
ellen.wei-rao@connect.polyu.hk, enmwmak@polyu.edu.hk

ABSTRACT

Conventional PLDA scoring in i-vector speaker verification involves the i-vectors of target speakers and claimants only. We have previously demonstrated that better performance can be achieved by incorporating the information of background speakers in the scoring process via speaker-dependent SVMs. This is achieved by defining a PLDA score space with dimension equal to the number of training i-vectors for each target speaker. The new protocol in NIST 2012 SRE permits systems to use the information of other target-speakers (called known non-targets) in each verification trial. In this paper, we exploit this new protocol to enhance the performance of PLDA-SVM scoring by using the score vectors of both known and unknown non-targets as the impostor class data to train the speaker-dependent SVMs. Because some target speakers have one enrollment utterance only, which results in severe imbalance in the speaker- and impostor-class data for SVM training. This paper shows that if the enrollment utterance is sufficiently long, a number of target-speaker i-vectors can be generated by an utterance partitioning and resampling technique, resulting in much better scoring SVMs. Results on NIST 2012 SRE demonstrate the advantages of pooling the known and unknown non-targets for training the SVMs and that the resampling techniques can help the SVM training algorithm to find better decision boundaries for those speakers with only a small number of enrollment utterances.

Index Terms— I-vectors; probabilistic linear discriminant analysis; empirical kernel maps; likelihood ratio kernels; NIST 2012 SRE.

1. INTRODUCTION

1.1. Motivation of Work

Current state-of-the-art speaker verification systems use i-vectors [1] as features and probabilistic linear discriminant analysis (PLDA) [2–4] as back-end classifiers. In the i-vector approach, the speaker and channel characteristics of an utterance are represented by the latent variables (factors) of a factor analyser [5,6] whose factor loading matrix – referred to as the total variability matrix – defines the subspace on which all i-vectors live. PLDA is then used for suppressing session variability in the i-vectors. In these systems, given a test i-vector and a target-speaker i-vector, the verification decision is based on the likelihood ratio (LR) score derived from two hypotheses: (1) the test i-vector and the target-speaker i-vector are from the same speaker and (2) these two i-vectors are from two different speakers. Because the computation of the likelihood ratio does not involve other i-vectors, this scoring method *implicitly* uses background information through the universal background model (UBM) [7] and the total variability matrix. This LR scoring method is computationally efficient, However, the implicit use of background information is a drawback of this method.

To address the limitation of PLDA scoring, we have recently proposed an empirical kernel SVM that takes the background speaker information *explicitly* during the scoring process [8]. This method captures the discrimination between a target-speaker and background-speakers in the SVM weights as well as in the score vectors that live in an empirical score space. Specifically, for each target speaker, an empirical score space with dimension equal to the number of training i-vectors for this target speaker is defined by using the idea of empirical kernel maps [9–11]. Given an i-vector, a score vector living in this space is formed by computing the LR scores of this i-vector with respect to each of the training i-vectors. A speaker-dependent support vector machine (SVM) – referred to as empirical LR SVM – can then be trained using the training score vectors. During verification, given a test i-vector and the target-speaker under test, the LR scores are mapped to a score vector, which is then fed to the target-speaker's SVM to obtain the final test score.

Compared to previous speaker recognition evaluations (SRE), NIST 2012 SRE [12] presents some new challenges to the research community, e.g., noise contaminated test segments and the severe variation in the length and number of enrolment utterances for target speakers. On the other hand, the evaluation also introduces some new protocols that open up opportunity for researchers to enhance system performance. In particular, the evaluation now allows systems to use the information of other target speakers for each verification trials. The permission to use other target speakers leads to the compound likelihood ratio [13–15] and anti-models [16], which improves verification performance substantially.

Unlike the compound likelihood ratio, this paper exploits the information of the known non-targets from another perspective. Specifically, instead of injecting the likelihood-ratio scores of known non-targets into the posterior probability computation as in [13, 15], this paper uses the PLDA scores arising from any i-vectors with respect to the target-speaker's i-vectors and a group of background speakers' i-vectors to define a speaker-dependent empirical score space. Then, for each target speaker, an SVM is trained by pooling the score vectors produced by all of the known and unknown non-targets.

SVM is one of the back-end classifiers adopted in the original i-vector approach [1]. However, SVM scoring is not very common in other i-vector systems, primary because of its inferior performance when compared with cosine distance scoring [1] and PLDA scoring [2]. The poorer performance of SVM scoring, however, is mainly due to the severe imbalance between the number of target-speaker

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

i-vectors and the number of background speaker i-vectors. Before NIST 2012 SRE, there is only one i-vector per target speaker, because there is only one enrollment session per target speaker. Although NIST 2012 SRE also provides multiple speech files for many target speakers, there are also target speakers who have one or a few enrollment sessions only. This difficulty, however, can be overcome by a technique called utterance partitioning with acoustic vector resampling (UP-AVR) [17, 18]. This technique has successfully boosted the performance of GMM-SVM [19–21] and i-vector based systems [22]. It has been demonstrated in [8, 22] that increasing the number of target-speaker i-vectors can help the SVM training algorithm to find better decision boundaries, thus making SVM scoring outperforms cosine-distance scoring and PLDA scoring.

1.2. Related Works

There has been previous work that uses known non-targets for SVM scoring in speaker verification. For example, [16] compares SVM scoring that uses unknown non-targets with SVM scoring that uses both unknown and known non-targets and shows the advantages of including the information of known non-targets for training SVMs. However, the way of using the known non-targets proposed in this paper is different from [16] in that the SVMs in this paper works on the empirical PLDA-LR score space, whereas the SVMs in [16] works on the GMM-supervector space.

2. EMPIRICAL LR KERNELS FOR SVMS

2.1. PLDA Likelihood-Ratio Scoring

Given a length-normalized [3] test i-vector $\mathbf{x}_t$ and target-speaker's i-vector $\mathbf{x}_s$, the likelihood ratio score can be computed as follows [3]:

$$\begin{aligned} S_{\text{LR}}(\mathbf{x}_t, \mathbf{x}_s) &= \frac{P(\mathbf{x}_t, \mathbf{x}_s|\text{same speaker})}{P(\mathbf{x}_t, \mathbf{x}_s|\text{different speakers})} \\ &= \frac{\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_t^\mathsf{T} & \mathbf{x}_s^\mathsf{T}\end{bmatrix}^\mathsf{T} \,|\, \begin{bmatrix}\boldsymbol{\mu}^\mathsf{T} & \boldsymbol{\mu}^\mathsf{T}\end{bmatrix}^\mathsf{T}, \tilde{\mathbf{V}}\tilde{\mathbf{V}}^\mathsf{T} + \tilde{\boldsymbol{\Sigma}}\right)}{\mathcal{N}\left(\begin{bmatrix}\mathbf{x}_t^\mathsf{T} & \mathbf{x}_s^\mathsf{T}\end{bmatrix}^\mathsf{T} \,|\, \begin{bmatrix}\boldsymbol{\mu}^\mathsf{T} & \boldsymbol{\mu}^\mathsf{T}\end{bmatrix}^\mathsf{T}, \text{diag}\{\mathbf{V}\mathbf{V}^\mathsf{T} + \boldsymbol{\Sigma}, \mathbf{V}\mathbf{V}^\mathsf{T} + \boldsymbol{\Sigma}\}\right)} \end{aligned} \tag{1}$$

where $\mathbf{V}$ is a factor loading matrix, $\boldsymbol{\mu}$ is the global mean of the i-vectors for training the PLDA model, $\boldsymbol{\Sigma}$ is full covariance matrix, $\tilde{\mathbf{V}} = \begin{bmatrix}\mathbf{V}^\mathsf{T} & \mathbf{V}^\mathsf{T}\end{bmatrix}^\mathsf{T}$ and $\tilde{\boldsymbol{\Sigma}} = \text{diag}\{\boldsymbol{\Sigma}, \boldsymbol{\Sigma}\}$. Details of PLDA scoring can be found in [2, 3, 8]. Using Eq. 1 and the standard formula for the inverse of block matrices [23], the log-likelihood ratio score is given by

$$S_{\text{LR}}(\mathbf{x}_t, \mathbf{x}_s) = \text{const} + \mathbf{x}_s^\mathsf{T}\mathbf{Q}\mathbf{x}_s + \mathbf{x}_t^\mathsf{T}\mathbf{Q}\mathbf{x}_t + 2\mathbf{x}_s^\mathsf{T}\mathbf{P}\mathbf{x}_t, \tag{2}$$

where

$$\begin{aligned} &\mathbf{P} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\Lambda} - \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})^{-1};\ \ \boldsymbol{\Lambda} = \mathbf{V}\mathbf{V}^\mathsf{T} + \boldsymbol{\Sigma} \\ &\mathbf{Q} = \boldsymbol{\Lambda}^{-1} - (\boldsymbol{\Lambda} - \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Gamma})^{-1};\ \ \boldsymbol{\Gamma} = \mathbf{V}\mathbf{V}^\mathsf{T}. \end{aligned} \tag{3}$$

2.2. Empirical Kernels and Empirical Kernel Maps

Eq. 2 and Eq. 3 suggest that PLDA LR scoring uses the information of background speakers implicitly. To make better use of the background information, we derived a speaker-dependent discriminative model for scoring called empirical LR SVM in [8].

Assuming that target-speaker s has H_s enrollment utterances, then H_s i-vectors will be obtained. In case the speaker provides one or a very small number of enrollment utterances only, we can apply an utterance partitioning technique [22] to produce multiple i-vectors from his/her enrollment utterance. Denote these i-vectors as:

$$\mathcal{X}_s = \{\mathbf{x}_{s,1}, \ldots, \mathbf{x}_{s,H_s}\}. \tag{4}$$

Let's denote the set of background-speaker i-vectors as:[1]

$$\mathcal{X}_b = \{\mathbf{x}_{b,1}, \ldots, \mathbf{x}_{b,B}\}. \tag{5}$$

Then, the SVM score of a test i-vector $\mathbf{x}_t$ is

$$\begin{aligned} S_{\text{SVM}}(\mathbf{x}_t, \mathcal{X}_s, \mathcal{X}_b) = &\sum_{j \in \text{SV}_s} \alpha_{s,j} K(\mathbf{x}_t, \mathbf{x}_{s,j}) - \\ &\sum_{j \in \text{SV}_b} \alpha_{s,j} K(\mathbf{x}_t, \mathbf{x}_{b,j}) + d_s \end{aligned} \tag{6}$$

where SV_s and SV_b contain the indexes of the support vectors corresponding to the speaker class and impostor class, respectively, and d_s is a speaker-dependent bias.

Here, we consider two possible kernels for $K(\cdot, \cdot)$, based on the idea of empirical kernel map [9–11].

Empirical LR Kernel I: Given a test i-vector $\mathbf{x}_t$,

$$K(\mathbf{x}_t, \mathbf{x}_{s,j}) = \mathbb{K}\left(\overrightarrow{S}_{\text{LR}}(\mathbf{x}_t, \mathcal{X}_s), \overrightarrow{S}_{\text{LR}}(\mathbf{x}_{s,j}, \mathcal{X}_s)\right) \tag{7}$$

where

$$\overrightarrow{S}_{\text{LR}}(\mathbf{x}_t, \mathcal{X}_s) = \begin{bmatrix} S_{\text{LR}}(\mathbf{x}_t, \mathbf{x}_{s,1}) \\ S_{\text{LR}}(\mathbf{x}_t, \mathbf{x}_{s,2}) \\ \vdots \\ S_{\text{LR}}(\mathbf{x}_t, \mathbf{x}_{s,H_s}) \end{bmatrix} \tag{8}$$

is an empirical kernel map, $S_{\text{LR}}(\mathbf{x}_t, \mathbf{x}_{s,i})$ is a PLDA score and $\mathbb{K}(\cdot, \cdot)$ is a standard SVM kernel, e.g., linear or RBF. Only RBF was adopted in this work. $\overrightarrow{S}_{\text{LR}}(\mathbf{x}_{s,j}, \mathcal{X}_s)$ can be obtained by replacing $\mathbf{x}_t$ in Eq. 8 with $\mathbf{x}_{s,j}$. Similar formulations apply to $K(\mathbf{x}_t, \mathbf{x}_{b,j})$ in Eq. 6. Note that the empirical feature space is defined by target-speaker's i-vectors through the PLDA model. Because H_s is typically small, the dimension of $\overrightarrow{S}_{\text{LR}}(\mathbf{x}_t, \mathcal{X}_s)$ is low. Therefore, it is possible to use a non-linear kernel for $\mathbb{K}(\cdot, \cdot)$.

Empirical LR Kernel II: Denote $\mathcal{X} = \{\mathcal{X}_s, \mathcal{X}_b\}$ as the training set for target-speaker s. Then,

$$K(\mathbf{x}_t, \mathbf{x}_{s,j}) = \mathbb{K}\left(\overrightarrow{S}_{\text{LR}}(\mathbf{x}_t, \mathcal{X}), \overrightarrow{S}_{\text{LR}}(\mathbf{x}_{s,j}, \mathcal{X})\right) \tag{9}$$

where

$$\overrightarrow{S}_{\text{LR}}(\mathbf{x}_t, \mathcal{X}) = \begin{bmatrix} S_{\text{LR}}(\mathbf{x}_t, \mathbf{x}_{s,1}) \\ \vdots \\ S_{\text{LR}}(\mathbf{x}_t, \mathbf{x}_{s,H_s}) \\ S_{\text{LR}}(\mathbf{x}_t, \mathbf{x}_{b,1}) \\ \vdots \\ S_{\text{LR}}(\mathbf{x}_t, \mathbf{x}_{b,B'}) \end{bmatrix} \tag{10}$$

where B' ($\leq B$) is the number of background i-vectors selected from the background speaker set $\mathcal{X}_b$. Unlike Eq. 8, the score vector in Eq. 10 also contains the LR scores of $\mathbf{x}_t$ with respect to the background i-vectors. As a result, discriminative information be-

[1] It is not necessary to apply partitioning to the utterances of background speakers because background i-vectors are abundant.

tween same-speaker pairs $\{\mathbf{x}_t, \mathbf{x}_{s,j}\}_{j=1}^{H_s}$ and different-speaker pairs $\{\mathbf{x}_t, \mathbf{x}_{b,j}\}_{j=1}^{B'}$ is embedded in the score vector. Note that the vector size in Eq. 10 is independent of the number of target-speakers. Therefore, the method is scalable to large systems with thousands of speakers.

3. KNOWN NON-TARGETS FOR EMPIRICAL LR SVM

In previous SREs, for each verification trial, only the knowledge of the target under test can be used for computing the score. This restriction, however, has been removed in NIST 2012 SRE. Therefore, known non-targets can be used to improve the discrimination power of empirical LR SVMs.

Assume that M target speakers have been enrolled in a system. When training the SVM of a target speaker, the remaining $(M-1)$ competing speakers from the target-speaker set are considered as the new background training set. Specifically, the speaker-class and impostor-class i-vectors for training the SVM of target speaker s are

$$\mathcal{X}_s = \{\mathbf{x}_{s,1}, \ldots, \mathbf{x}_{s,H_s}\} \text{ and } \mathcal{X}_a = \{\mathbf{x}_{a,1}, \ldots, \mathbf{x}_{a,M-1}\}, \quad (11)$$

respectively, where $\mathcal{X}_a$ contains the i-vectors of the competing known non-targets with respect to s. As a result, the SVM score of a test i-vector $\mathbf{x}_t$ is

$$S'_{\text{SVM}}(\mathbf{x}_t, \mathcal{X}_s, \mathcal{X}_a) = \sum_{j \in \text{SV}_s} \alpha_{s,j} K(\mathbf{x}_t, \mathbf{x}_{s,j}) - \sum_{j \in \text{SV}_a} \alpha_{s,j} K(\mathbf{x}_t, \mathbf{x}_{a,j}) + d'_s \quad (12)$$

where $K(\cdot,\cdot)$ is an empirical LR kernel (Eq. 7 or Eq. 9), SV_s and SV_a contain the indexes of the support vectors corresponding to the speaker class and impostor class, respectively, and d'_s is a speaker-dependent bias.

4. EXPERIMENTS AND RESULTS

4.1. Speech Data and Acoustic Features

The *core set* of NIST 2012 Speaker Recognition Evaluation (SRE) [12] was used for performance evaluation. This paper focuses on the male phone-call speech of the core task, i.e., Common Evaluation Conditions 2, 4, and 5. In the evaluation dataset, no noise was added to the test segments of common condition 2, whereas noise was added to the test segments of common condition 4 and test segments in common condition 5 were collected in a noisy environment. All of these conditions contain training segments with variable length and variable numbers of training segments per target speaker. We removed the 10-second utterances and the summed-channel utterances from the training segments of NIST 2012 SRE but ensured that all target speakers have at least one long utterance for training. In the sequel, we use "CC" to denote common evaluation conditions. The speech files of male speakers in NIST 2005–2010 SREs were used as development data for training the UBM, total variability matrix, LDA-WCCN, PLDA models, and Z-Norm parameters [24].

We used our recently proposed voice activity detector [25, 26] to detect the speech regions of each utterance. 19 MFCCs together with energy plus their 1st- and 2nd- derivatives were extracted from the speech regions, followed by cepstral mean normalization [27] and feature warping [28] with a window size of 3 seconds. A 60-dim acoustic vector was extracted every 10ms, using a Hamming window of 25ms.

To improve the noise robustness, we followed the suggestions in [14] to add noise to the training files. To this end, we constructed a noise dataset comprising 13 real crowd noise files and 17 heating, ventilation, and air conditioning (HVAC) noise files from [29] and 10 artificial crowd noise files generated by summing 441 utterances from male and female speakers in pre-2012 NIST SRE. For each training file with SNR above 15dB, we generated two noisy speech files at an SNR of 6dB and 15dB by randomly selecting two noise files from the noise dataset. For each training file with SNR between 6dB and 15dB, we produced a noisy speech file at 6dB.

4.2. Total Variability Modeling and PLDA

The i-vector systems are based on a gender-dependent UBM with 1024 mixtures. 3,500 microphone utterances and 3,501 telephone utterances from NIST 2005–2008 SREs were used for training the UBM. We selected 14,875 telephone and interview conversations from 575 speakers in NIST 2006–2010 SREs to estimate a total variability matrix with 400 total factors.

According to [30], adding noise to the training files of UBM and total variability modeling receives insignificant performance gain. Hence, we followed the steps in [30] and only added noise to the training files of LDA and PLDA models. For the common condition without added noise (CC2), we selected 15,662 original utterances from 673 male speakers from NIST 2006-2010 SREs to estimate the loading matrix of Gaussian PLDA. For the common conditions that comprise noisy test segments (CC4 and CC5), we pooled 15,662 original utterances, 14,353 utterances at 6 dB SNR, and 10,932 utterances at 15 dB SNR to estimate the loading matrix.

We applied whitening [31] and i-vector length normalization [3] to the 400-dimensional i-vectors. Then, we performed linear discriminant analysis (LDA) [32] and within-class covariance normalization (WCCN) [31] on the resulting vectors to reduce the dimension to 200 before training the PLDA models with 150 latent variables.

4.3. The Effect of Using Noisy Training Files

Table 2 shows the effect of adding noise to the training utterances on the PLDA models for the common conditions involving noisy test segments. We pooled the original i-vectors, 6 dB i-vectors, and 15 dB i-vectors for enrollment and for training the PLDA models. According to the results in Table 2, the strategy of adding noise to the training speech files can boost the performance of i-vector based PLDA systems, especially for CC4.

Method	EER (%)		MinNDCF	
	CC4	CC5	CC4	CC5
PLDA	4.35	**2.74**	0.43	0.36
PLDA with Added Noise	**3.00**	3.23	**0.33**	**0.34**

Table 2. Performance of PLDA scoring in common conditions 4 and 5 of NIST 2012 SRE. *PLDA*: original speech files were used for enrollment and for training the PLDA model. *PLDA with Added Noise*: HAVC and crowd noises were added to the enrollment and PLDA training files at 6dB and 15dB SNR.

4.4. Known Non-targets versus Unknown Non-targets

We considered the classical LR scoring based on Gaussian PLDA with added noise as the baseline (PLDA in Table 1). Because no

	Method	Source of Imposter Class for Training SVMs	EER (%)			MinNDCF		
			CC2	CC4	CC5	CC2	CC4	CC5
1	PLDA	–	2.40	3.00	3.23	0.33	0.33	0.34
2	PLDA+UP-AVR	–	2.32	3.13	3.32	0.32	0.31	0.33
3	PLDA+UP-AVR+SVM-I	Unknown non-targets	1.94	2.80	2.72	0.31	0.30	0.33
4	PLDA+UP-AVR+SVM-I	Known non-targets	1.93	2.70	2.71	0.32	0.30	0.33
5	PLDA+UP-AVR+SVM-I	Known + Unknown non-targets	1.90	2.70	2.73	0.31	0.30	0.32
6	PLDA+SVM-II	Unknown non-targets	2.15	3.16	2.71	0.32	0.28	0.31
7	PLDA+UP-AVR+SVM-II	Unknown non-targets	1.94	2.77	2.59	0.34	0.29	0.31
8	PLDA+UP-AVR+SVM-II	Known non-targets	**1.84**	**2.69**	2.61	**0.31**	0.29	0.32
9	PLDA+UP-AVR+SVM-II	Known + Unknown non-targets	1.84	2.70	**2.54**	0.31	**0.28**	**0.31**

Table 1. Performance of various scoring methods for NIST 2012 SRE (male speakers) under the common conditions that involve telephone recordings. The methods are named by the processes applied to the i-vectors for computing the verification scores. For example, *PLDA+UP-AVR+SVM-II* means that UP-AVR has been applied to create target-speaker i-vectors for training SVMs that use Empirical LR Kernel II (Eq. 9 and Eq. 10). Note that because some target speakers have one enrollment utterance only, it is impossible to apply empirical LR kernel I without UP-AVR. Therefore, no results for PLDA+SVM-I are reported.

noise was added to the test segments of CC2 while those in CC4 and CC5 are noisy, different experimental setups were applied to these three common conditions. For CC2, we selected 704 unknown non-targets or 722 known non-targets as the background speakers, i.e., $B = 704$ or $B = 722$ in Eq. 5. We used the i-vectors of these speakers as the impostor-class data to train an SVM for each target-speaker using the empirical LR kernels described in Section 2.2. The penalty factor was set to 1.0 for all SVMs and parameter in the RBF kernel $\mathbb{K}$ was set to 40. For the ease of finding this parameter, we performed Z-norm on the PLDA LR scores and used the normalized scores to construct the empirical kernels.

Although many target speakers in NIST SRE 2012 have multiple training segments, some of them have a few training segments only. More precisely, after removing the 10-second segments and summed-channel segments , 50 out of 723 target speakers have one long training segments only and 390 of them have less than 17 training segments. Therefore, UP-AVR was applied to alleviate the effect of data imbalance in training the SVMs whenever the number of training segments per target speaker is less than 17. The number of partitions per enrollment utterance and the number of resampling in UP-AVR were set to 4. To reduce scoring time, B' in Eq. 10 was set to 150. The settings of UP-AVR in Table 1 are identical for all systems.

For CC4 and CC5, because the test segments contain noise, we also prepared B ($B = 704$ or $B = 722$) imposter speaker utterances at 6 dB SNR and 15 dB SNR for condition matching and used the i-vectors of these $3B$ imposter utterances to train the SVMs. This setting was applied to both SVM-I and SVM-II in Table 1. In addition, the parameter in the RBF kernel $\mathbb{K}$ was set to 90. The other settings of CC4 and CC5 are the same as CC2.

Row 3 and Row 4 in Table 1 suggest that the effect of including known non-targets on SVM-I is small, but the known non-targets can still reduce the EER. For SVM-II, Row 7 and Row 8 in Table 1 demonstrate the advantage of using known non-targets over using unknown non-targets. The results in Rows 7 to 9 demonstrate the advantages of pooling the known and unknown non-targets for training the SVMs.

4.5. UP-AVR for SVM Scoring and LR Scoring

Rows 6 and 7 suggest that UP-AVR is very important for SVM scoring. After applying UP-AVR, the performance of SVMs scoring improves significantly and is much better than PLDA scoring. UP-AVR not only helps to alleviate the data-imbalance problem in SVM training, but also enriches the information content of the scoring vectors by increasing the number of LR scores derived from the target speaker. However, UP-AVR is not beneficial to LR scoring, as evident by the performance of *PLDA* and *PLDA+UP-AVR* in Table 1.

5. CONCLUSION

Inspired by the new challenges and protocols in NIST 2012 SRE, this paper takes the advantage of empirical kernel maps and utilizes the information of known non-targets to train SVMs with high discriminative power. In addition, this paper introduces utterance partitioning with acoustic vector resampling to maximizing the utilization of speech data and mitigate the data-imbalance problem on training SVMs. Results on NIST 2012 SRE show the advantages of using known non-targets comparing with unknown non-targets and suggest that the idea of incorporating known non-targets information into the training of speaker-dependent SVMs together with the utterance partitioning techniques can boost the performance of i-vector based PLDA systems significantly. The idea of incorporating known non-targets can be further explored in future work. For example, the known non-targets can also be used for defining the score space.

6. REFERENCES

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

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

[3] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems,"

in *Proc. of Interspeech 2011*, Florence, Italy, Aug. 2011, pp. 249–252.

[4] S.J.D. Prince and J.H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007.

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

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

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, Jan. 2000.

[8] M.W. Mak and W. Rao, "Likelihood-ratio empirical kernels for i-vector based PLDA-SVM scoring," in *Proc. ICASSP 2013*, Vancouver, Canada, May 2013, pp. 7702–7706.

[9] B. Scholkopf, S. Mika, C. J. C. Burges, P. Knirsch, K. R. Muller, G. Ratsch, and A. J. Smola, "Input space versus feature space in kernel-based methods," *IEEE Trans. on Neural Networks*, vol. 10, no. 5, pp. 1000–1017, Sept. 1999.

[10] H. Xiong, M.N.S Swamy, and M.O. Ahmad, "Optimizing the kernel in the empirical feature space," *IEEE Trans. on Neural Networks*, vol. 16, no. 2, pp. 460 – 474, 2005.

[11] S. X. Zhang and M. W. Mak, "Optimized Discriminative Kernel for SVM Scoring and Its Application to Speaker Verification," *IEEE Trans. on Neural Networks*, vol. 22, no. 2, pp. 173–185, 2011.

[12] NIST 2012 speaker recognition evaluation, "http://www.nist.gov/itl/iad/mig/sre12results.cfm," 2012.

[13] N. Brummer, "LLR transformation for SRE'12," Dec. 2012, Notes relevant to the NIST 2012 SRE, Online: https://sites.google.com/site/bosaristoolkit/sre12.

[14] D. A.van Leeuwen and R. Saeidi, "Knowing the non-target speakers: The effect of the i-vector population for PLDA training in speaker recognition," in *Proc. ICASSP 2013*, Vancouver, BC, Canada, May 2013, pp. 6778 – 6782.

[15] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition system," in *Proc. ICASSP 2013*, Vancouver, BC, Canada, May 2013, pp. 7663 – 7667.

[16] H. W. Sun, K. A. Lee, and B. Ma, "Anti-model KL-SVM-NAP system for NIST SRE 2012 evaluation," in *Proc. ICASSP 2013*, Vancouver, BC, Canada, May 2013, pp. 7688 – 7692.

[17] W. Rao and M. W. Mak, "Addressing the data-imbalance problem in kernel-based speaker verification via utterance partitioning and speaker comparison," in *Proc. of Interspeech 2011*, Florence, Aug. 2011, pp. 2717–2720.

[18] M. W. Mak and W. Rao, "Utterance partitioning with acoustic vector resampling for GMM-SVM speaker verification," *Speech Communication*, vol. 53, no. 1, pp. 119–130, Jan. 2011.

[19] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[20] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, Toulouse, France, May 2006, vol. 1, pp. 97–100.

[21] A. Solomonoff, C. Quillen, and W. M. Campbell, "Speaker verification using support vector machines and high-level features," *IEEE Transactions On Audio, Speech, and Language Processing*, vol. 15, no. 7, SEPTEMBER 2007.

[22] W. Rao and M. W. Mak, "Boosting the performance of i-vector based speaker verification via utterance partitioning," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 21, no. 5, pp. 1012 – 1022, 2013.

[23] K. B. Petersen and M. S. Pedersen, *The matrix cookbook*, [Online]. Available: http://matrixcookbook.com, 2008.

[24] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 42–54, Jan. 2000.

[25] H.B. Yu and M.W. Mak, "Comparison of voice activity detectors for interview speech in NIST speaker recognition evaluation," in *Proc. of Interspeech 2011*, Florence, Aug. 2011, pp. 2353–2356.

[26] M.W. Mak and H.B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations," *Computer Speech and Language*, vol. 28, no. 1, pp. 295 – 313, Jan. 2014.

[27] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, Jun. 1974.

[28] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. of Odyssey: Speaker and Language Recognition Workshop*, Crete, Greece, Jun. 2001, pp. 213–218.

[29] [Online]. Available: www.freesound.org.

[30] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. ICASSP 2012*, Kyoto, Japan, March 2012, pp. 4253 – 4256.

[31] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. of the 9th International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, Sep. 2006, pp. 1471–1474.

[32] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.