論文 / 著書情報
Article / Book Information

Title	Autonomous Selection of i-Vectors for PLDA Modelling in Speaker Verification
Authors	Sangeeta Biswas, Johan Rohdin, Koichi Shinoda
Citation	Elsevier Speech Communication, vol. 72, , pp. 32-46
Pub. date	2015, 5
DOI	https://doi.org/10.1016/j.specom.2015.05.001
Creative Commons	See next page.
Note	このファイルは著者（最終）版です。 This file is author (final) version.

License

Autonomous Selection of i-Vectors for PLDA Modelling in Speaker Verification

Sangeeta Biswas, Johan Rohdin, and Koichi Shinoda

Department of Computer Science
Tokyo Institute of Technology, Japan

Abstract

Recently, systems combining i-vector and probabilistic linear discriminant analysis (PLDA) have become one of the state-of-the-art methods in text-independent speaker verification. The training data of a PLDA model is often collected from a large, diverse population. However, including irrelevant or noisy training data may deteriorate the verification performance. In this paper, we first show that data selection using k-NN improves the speaker verification performance. We then present a robust way of selecting k based on the *local distance-based outlier factor* (LDOF). We call this method *flexible* k-NN (fk-NN). We conduct experiments on male and female trials of several telephone conditions of the NIST 2006, 2008, 2010 and 2012 Speaker Recognition Evaluations (SRE). By using fk-NN, we discard a substantial amount of irrelevant or noisy training data without depending on tuning k, and achieve significant performance improvements on the NIST SRE sets.

1. Introduction

Data selection is an important issue in speaker recognition. In previous studies, data selection for the impostor models in T-normalisation (Sturim & Reynolds, 2005; McLaren et al., 2009), for the background dataset of support vector machines (SVM) (McLaren et al., 2010; Suh et al., 2011), or for a universal background model (UBM) (Hasan et al., 2010; Hasan & Hansen, 2011; Huang & Ma, 2011) were addressed. It was shown that data relevancy is more important for the verification performance than data size. In this paper, we address the relevant data selection issue for *probabilistic linear discriminant analysis* (PLDA) (Ioffe, 2006; Prince & Elder, 2007) modelling in i-vector based text-independent speaker verification.

An i-vector system maps an utterance into a low dimensional subspace, known as the *total variability subspace* (Dehak et al., 2009, 2011). The coordinate vector in the total variability subspace is known as an i-vector. An i-vector contains information related to an speaker identity, as well as irrelevant factors such as the transmission channels or the speaker's emotion. Currently, PLDA is one of the state-of-the-art methods for separating the speaker identity from irrelevant factors and generating a likelihood-ratio score for two given i-vectors.

In order to train the parameters of a PLDA model, multi-session recordings from several hundred speakers, resulting in several thousands of recordings from multiple databases, are typically used. For example, research groups involved in the NIST speaker recognition evaluation (SRE) typically use utterances from all NIST 2004-2005 data along with the Switchboard II, Phases 1, 2 and 3; Switchboard Cellular, Parts 1 and 2 data, and Fisher data. However, there is no evidence that using all the available data would guarantee the best PLDA model.

Based on the experiences from the other models such as UBM, SVM or joint factor analysis (JFA), researchers typically use gender-dependent PLDA models. Senoussaoui et al. (2011) empirically showed that gender-dependent PLDA models outperformed gender-independent PLDA models. Kanagasundaram et al. (2012) showed that the PLDA model trained by utterances whose lengths matched with those utterances in the evaluation set performed better than that trained by full-length utterances. These studies indicate that in order to get better performance from a PLDA model, it is necessary to ensure that the training data of the PLDA model matches the properties of the target evaluation set. However, it is not always obvious which properties are important to be matched. Therefore, in Biswas et al. (2014), a data-driven approach was adopted. This paper is an extended version of Biswas et al. (2014).

In many applications such as on-line bank services for registered customers, we can access the set of speakers enrolled to the system, i.e., *enrolment set*, during the development phase of the system. Targeting such applications, we proposed to use the enrolment set for selecting suitable training data for the PLDA model in Biswas et al. (2014). We showed that by selecting a training set whose i-vectors are close to the i-vectors of the enrolment set, the PLDA modelling can be improved. We first used the k-NN method in order to choose the k-nearest neighbours of each enrolment speaker in the training set of the PLDA model. We showed that this method performs remarkably well when the optimal k is known. However, it is difficult to estimate the optimal k. We, therefore, proposed a robust way of selecting k based on the *local distance-based outlier factor* (LDOF) (Zhang et al., 2009). We named our method *flexible* k-NN (fk-NN).

In Biswas et al. (2014), we conducted experiments using data without any noise for the PLDA modelling. In our experiments, we used the same k for all enrolment speakers in order to avoid complexities. In this paper, we deal with more realistic scenarios, where the training set of the PLDA model and the authentication set[1] are noisy. We also discuss how to use the enrolment speaker dependent k instead of using the same k for all enrolment speakers in fk-NN. The effect of i-vector selec-

[1] A set of test segments

tion on known and unknown non-target trials, is also a topic of this paper.

We evaluate the data selection methods on the NIST SRE 2006 core task, the NIST SRE 2008 core task (condition-6), the NIST SRE 2010 extended core task (condition-5), and the telephone conditions of the NIST SRE 2012 core task (condition-2, -4 and -5). All these evaluation sets have only telephone data in the authentication sets. Note that using the knowledge of the enrolment set for system development is allowed in the NIST SRE 2012 but not in the earlier NIST SREs. Our experiments show that fk-NN can obtain significant performance improvements on both male and female trials by discarding a substantial amount of irrelevant or noisy training data of the PLDA model.

Since the training time of the PLDA model is very short (several seconds), fk-NN does not offer a large reduction in computational expense as some data selection methods for UBM training (Hasan & Hansen, 2011) do. However, it improves the verification accuracy. Thanks to the short training time, we can re-train the PLDA model quickly after adding relevant data for newly added enrolment speakers, which would not be practical for UBM training or other offline modelling such as factor loading matrix (e.g., total variability matrix) training.

The organisation of this paper is as follows: Section 2 introduces the i-vector based speaker verification system, Section 3 describes PLDA modelling for i-vectors, Section 4 presents our data selection methods, and Section 5 experimentally evaluates the effect of selecting i-vectors for PLDA modelling. Finally, Section 6 draws conclusions of this paper.

2. i-Vector based speaker verification

An i-vector based system (Dehak et al., 2009, 2011) assumes that the feature vectors of an utterance are drawn independently from a Gaussian mixture model (GMM). The stacked mean vectors of the GMM constitute a speaker- and channel-dependent *GMM-supervector*, $\boldsymbol{\mu}$. It is assumed that $\boldsymbol{\mu}$ is generated according to

$$\boldsymbol{\mu} = \bar{\boldsymbol{\mu}} + \mathbf{T}\boldsymbol{\phi}, \tag{1}$$

where $\bar{\boldsymbol{\mu}}$ is the mean of speaker- and channel-independent supervectors, $\mathbf{T}$ is a basis for the *total variability subspace*, and $\boldsymbol{\phi}$ is a random vector. It is assumed that $\boldsymbol{\phi}$ follows the standard normal distribution and its dimension is lower than that of $\bar{\boldsymbol{\mu}}$.

Given the features from an utterance, the i-vector, $\boldsymbol{\omega}$, is the *maximum a posteriori* (MAP) estimate of $\boldsymbol{\phi}$. The mathematical framework for training $\mathbf{T}$ and estimating $\boldsymbol{\phi}$ is the same as used for training the eigenvoice matrix, $\mathbf{V}$, and estimating the hidden variable, $\mathbf{y}$, in the eigenvoice MAP (Kenny et al., 2005). The only difference is that, in the eigenvoice MAP, $\mathbf{y}$ is the same for all utterances of the same speaker, whereas in an i-vector based system, $\boldsymbol{\phi}$ is different from utterance to utterance.

An i-vector contains information not only about the speaker identity but also to a large extent about other factors such as the speaker's emotions, transmission channels, languages, and environmental noises. These other factors in all can be referred to as *channel* factors and should ideally be removed before verification. Three popular channel compensation techniques, namely within class covariance normalisation (WCCN) (Hatch et al., 2006), linear discriminate analysis (LDA), and nuisance attribute projection (NAP) (Campbell et al., 2006), were used to remove the effect of channel factors from the i-vectors in Dehak et al. (2011). The low dimension of the i-vector inspired researchers to use more advanced methods. Currently, PLDA introduced in Kenny (2010) has became one of the state-of-the-art methods for removing channel effects from i-vectors in text-independent speaker verification.

3. PLDA modelling

PLDA was originally proposed for object recognition in image processing independently by Ioffe (2006) and Prince & Elder (2007). Prince & Elder (2007) assumed that the feature vector, $\mathbf{g}$, is generated as:

$$\mathbf{g} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{m}$ is the mean of $\mathbf{g}$, and $\mathbf{y}$ and $\mathbf{x}$ are random vectors dependent on the class and channel factors, respectively. The vector $\boldsymbol{\epsilon}$ also depends on the channel factors and follows $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal covariance matrix. The vectors $\mathbf{y}$ and $\mathbf{x}$ follow the standard normal distribution. The matrix $\mathbf{V}$ is a basis for the *between-class subspace* and the matrix $\mathbf{U}$ is a basis for the *within-class subspace*. This PLDA model is very similar to the joint factor analysis (JFA) model (Kenny, 2005; Kenny et al., 2007) proposed for speaker recognition using the GMM-supervector, $\boldsymbol{\mu}$. The difference is that in the PLDA model, $\mathbf{g}$ is observed whereas in JFA, $\boldsymbol{\mu}$ is *indirectly observed*, i.e., we observe features drawn from the GMM but we do not know the parameters of $\boldsymbol{\mu}$.

Kenny introduced PLDA as in Eq. (2) for speaker verification with i-vectors as features in Kenny (2010). The author suggested to skip $\mathbf{U}\mathbf{x}$ but instead use full covariance $\boldsymbol{\Sigma}$ when large amounts of data are available, i.e.,

$$\boldsymbol{\omega} = \mathbf{m} + \mathbf{V}\mathbf{y} + \boldsymbol{\epsilon}. \tag{3}$$

Using a full covariance matrix, $\boldsymbol{\Sigma}$, is possible since the dimension of the i-vector is low. The PLDA model in Eq. (3) is similar to the *two-covariance model* proposed by Brümmer & Villiers (2010) and to the PLDA model proposed by Ioffe (2006). The rank of $\mathbf{V}$ is lower than the dimension of the feature vector in Ioffe (2006); Kenny (2010). On the other hand, in Brümmer & Villiers (2010), the rank of $\mathbf{V}$ is equal to the dimension of the feature vector, which means that the between-class covariance $\mathbf{V}\mathbf{V}^T$ has a full rank.

Given the two i-vectors, $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_j$ involved in a trial, the verification score, s_{ij}, is computed as:

$$s_{ij} = \log \frac{p(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j|\mathcal{H}_s)}{p(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j|\mathcal{H}_d)}, \tag{4}$$

where $\mathcal{H}_s$ and $\mathcal{H}_d$ are the following two hypotheses
$\mathcal{H}_s$: $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_j$ belong to the same speaker
$\mathcal{H}_d$: $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_j$ belong to two different speakers

When $\mathbf{m} = \mathbf{0}$, the closed-form solution of Eq. (4) is

$$s_{ij} = 2\boldsymbol{\omega}_i^T \mathbf{P} \boldsymbol{\omega}_j + \boldsymbol{\omega}_i^T \mathbf{Q} \boldsymbol{\omega}_i + \boldsymbol{\omega}_j^T \mathbf{Q} \boldsymbol{\omega}_j + c, \quad (5)$$

where c is a constant, and

$$\mathbf{P} = \boldsymbol{\Sigma}_a^{-1} \boldsymbol{\Sigma}_b - (\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_b \boldsymbol{\Sigma}_a^{-1} \boldsymbol{\Sigma}_b)^{-1}, \quad (6)$$

$$\mathbf{Q} = \boldsymbol{\Sigma}_a^{-1} - (\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_b \boldsymbol{\Sigma}_a^{-1} \boldsymbol{\Sigma}_b)^{-1}, \quad (7)$$

where $\boldsymbol{\Sigma}_a = \mathbf{V}\mathbf{V}^T + \boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}_b = \mathbf{V}\mathbf{V}^T$ [see Garcia-Romero & Espy-Wilson (2011)].

Typically, $\mathbf{V}$ and $\boldsymbol{\Sigma}$ are estimated by maximising the likelihood (ML) of the training data by means of an EM algorithm (Prince & Elder, 2007). It has been shown in Garcia-Romero & Espy-Wilson (2011) that it is better to apply whitening followed by length normalisation to the i-vectors before estimating the parameters of the PLDA model in order to make the i-vectors more closely follow a Gaussian distribution.

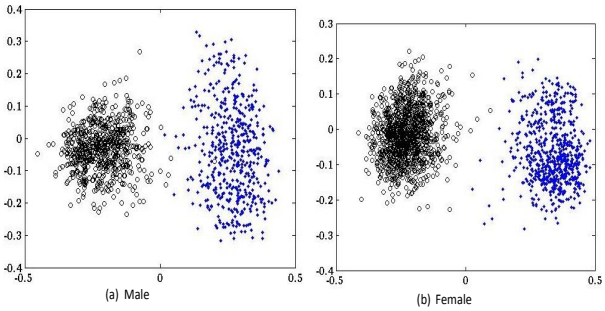

Figure 1: Clusters formed by i-vectors extracted from phone- and microphone-utterances after principal component analysis (PCA). x-axis is for the first principal component and y-axis is for the second principal component. Black circles represent i-vectors extracted from the phone-utterances and blue dots represent i-vectors extracted from the microphone-utterances. (a) There are 1270 male i-vectors. Among them 648 are from the phone-utterances and 622 are from the microphone-utterances. (b) There are 1993 female i-vectors. Among them 1140 are from the phone-utterances and 853 are from the microphone-utterances.

4. i-Vector selection

4.1. Overview

Let the sets of i-vectors, $\boldsymbol{\omega}$, for the PLDA modelling, for enrolment speakers, and for authentication be $\mathcal{P}$, $\mathcal{E}$, and $\mathcal{A}$, respectively. In order to train a good PLDA model, two conditions need to be fulfilled. First, $\mathcal{P}$ should be plentiful. Multi-session recordings from several hundred speakers, resulting in several thousands of recordings are typically needed. Second, $\mathcal{P}$ should be relevant; $\mathcal{P}$ should have similar properties to $\mathcal{E}$ and $\mathcal{A}$. There is a trade-off between these two conditions. Gender-dependent $\mathcal{P}$ is one good compromise for this trade-off. Senoussaoui et al. (2011) empirically showed that a gender-dependent PLDA model outperformed a gender-independent one. Obviously, a speaker's acoustic properties depend not only on gender but also on the physical properties of the vocal tract, dialect, age etc. In addition, phone sets, transmission channel types or background noises are known to greatly affect the acoustic properties of a recording. Kanagasundaram et al. (2012) showed that when $\boldsymbol{\omega}_p \in \mathcal{P}$ were extracted from utterances whose lengths matched with utterances used for $\boldsymbol{\omega}_e \in \mathcal{E}$ and $\boldsymbol{\omega}_a \in \mathcal{A}$, an improvement was achieved over the PLDA model trained by $\boldsymbol{\omega}_p \in \mathcal{P}$ extracted from full-length utterances. It seems therefore natural to select the training data based on more properties than gender. Since we do not know what other properties are important to consider, in this paper we adopt a data-driven approach.

In our approach, we target the application where we can access $\mathcal{E}$ during the development phase of the system. Given $\mathcal{E}$, we try to find a training set, $\mathcal{S} \subset \mathcal{P}$, that has similar properties to $\mathcal{E}$. In general, i-vectors having similar properties are close to each other. One scenario is visualised in Fig. 1 where microphone and telephone recordings are clearly separated[2]. We can safely assume that the set of $\boldsymbol{\omega}_p \in \mathcal{P}$ that has smaller distance from the set of $\boldsymbol{\omega}_e \in \mathcal{E}$, can be our desired $\mathcal{S}$.

[2] In our experiments, we are aiming at finding $\mathcal{S}$ among telephone recordings only.

In this study, we apply k-nearest neighbour (k-NN) in order to find $\mathcal{S}$. We show that this method performs remarkably well when the optimum k is known. However, it is difficult to estimate the optimum k. The optimum k may vary between different $\mathcal{E}$. We, therefore, present a robust way of selecting k based on the *local distance-based outlier factor* (LDOF) (Zhang et al., 2009).

4.2. k-NN

Let $\mathcal{S}_e^k \subset \mathcal{P}$ be the set of the k-nearest neighbours of an enrolment i-vector, $\boldsymbol{\omega}_e$. The steps of our data selection process using k-NN are given as:

1. Set the value of k.
2. For each $\boldsymbol{\omega}_e \in \mathcal{E}$, find $\mathcal{S}_e^k$.
   (a) Estimate the distance from $\boldsymbol{\omega}_e$ to each $\boldsymbol{\omega}_p \in P$, i.e., $\text{dist}(\boldsymbol{\omega}_e, \boldsymbol{\omega}_p)$.
   (b) Sort $\text{dist}(\boldsymbol{\omega}_e, \boldsymbol{\omega}_p)$ in ascending order.
   (c) Put the k-nearest neighbours of $\boldsymbol{\omega}_e$ from the set of $\boldsymbol{\omega}_p \in \mathcal{P}$ into $\mathcal{S}_e^k$.
3. Take the unique set of i-vectors from $\{\mathcal{S}_e^k\}_{\forall e}$ to get $\mathcal{S}$.

We can choose the value of k from a range of values for which we can get the best verification accuracy on a development set. However, selecting the optimal value of k based on the verification accuracy, to some extent depends on the range of k's values and the step-size. If we choose small range and large step-size, we would not get the optimum value. On the other hand, a large range and a small step size cause a computationally expensive k optimisation process. The second problem is that k may vary from database to database. Therefore, one k does not guarantee good result in all evaluation sets. Furthermore, the size and the spreadness of $\mathcal{E}$ compared to the spreadness of $\mathcal{P}$ may affect the number of selected i-vectors. If the i-vectors in $\mathcal{E}$ are close to each other compared to the typical distance between the i-vectors in $\mathcal{P}$, then every $\boldsymbol{\omega}_e \in \mathcal{E}$ will select almost the same $\boldsymbol{\omega}_p \in \mathcal{P}$. In such case, if the size of $\mathcal{E}$ is very small, we need a large k in order to get a sufficient amount

of i-vectors for training a good PLDA model. On the other hand, if the size of $\mathcal{E}$ is large, then a large k may select unnecessary data. Therefore, if we use a k optimised for a different $\mathcal{E}$, we may not get a sufficient amount of i-vectors for training a good PLDA model, or we may end up covering almost the whole training set, $\mathcal{P}$. Another more complicated problem is that the i-vectors in $\mathcal{S}_e^k$ might be much closer to each other than they are to $\boldsymbol{\omega}_e \in \mathcal{E}$. In this case, the i-vectors in $\mathcal{S}_e^k$ form a cluster, and $\boldsymbol{\omega}_e$ becomes its *outlier*. In such a case, $\mathcal{S}_e^k$ cannot be expected to improve modelling of the region surrounding $\boldsymbol{\omega}_e$.

In order to solve these problems, we have proposed a modification of the k-NN method, which we have named *flexible* k-NN (fk-NN) (Biswas et al., 2014). In this method, we first use the LDOF defined in the next section to measure to what extent $\boldsymbol{\omega}_e$ deviates from the cluster made by $\mathcal{S}_e^k$. We then increase k until all $\boldsymbol{\omega}_e \in \mathcal{E}$ lie inside the cloud of nearest neighbours according to the LDOF criteria. Our proposed fk-NN helps to decide k based on the nature of the target $\mathcal{E}$, not on any development set.

4.3. Flexible k-NN (fk-NN)

4.3.1. LDOF

In data mining applications, LDOF proposed by Zhang et al. (2009) is used for capturing the outlierness of an object among a scattered neighbourhood. In this paper, we use it to control the value of k in the k-NN based data selection process. The LDOF of $\boldsymbol{\omega}_e$ given k is defined as

$$\text{LDOF}_e^k = \frac{d_e^k}{D_e^k}, \tag{8}$$

where d_e^k is the k-NN distance of $\boldsymbol{\omega}_e$ and D_e^k is the k-NN inner distance of $\boldsymbol{\omega}_e$'s neighbourhood, which are defined as:

$$d_e^k = \frac{1}{k} \sum_{\boldsymbol{\omega}_i \in \mathcal{S}_e^k} \text{dist}(\boldsymbol{\omega}_e, \boldsymbol{\omega}_i), \tag{9}$$

$$D_e^k = \frac{1}{k(k-1)} \sum_{\boldsymbol{\omega}_i, \boldsymbol{\omega}_j \in \mathcal{S}_e^k, i \neq j} \text{dist}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j). \tag{10}$$

As shown in Fig. 2, LDOF captures the degree to which $\boldsymbol{\omega}_e$ deviates from its neighbourhood $\mathcal{S}_e^k$. When $\text{LDOF}_e^k \leq 1$, we can say that $\boldsymbol{\omega}_e$ is surrounded by the cloud created by the i-vectors of $\mathcal{S}_e^k$. Notice that if $k = 1$, D_e^k is undefined, therefore, LDOF cannot be calculated.

4.3.2. Algorithm of fk-NN

The most naive way of applying LDOF for determining k in k-NN is to select an individual value of k, k_e, for each $\boldsymbol{\omega}_e$, as the minimum from among those k's which suffice $\text{LDOF}_e^k \leq 1$. However, LDOF is not robustly estimated for values of k much smaller than the dimension of the data. In order to use LDOF in a cautious way, we therefore first find the minimum k which suffice $\text{LDOF}_e^k \leq 1$ for all $\boldsymbol{\omega}_e$. We then use this k for all $\boldsymbol{\omega}_e$. We call this method fk-NN (Biswas et al., 2014). The fk-NN algorithm is as follows:

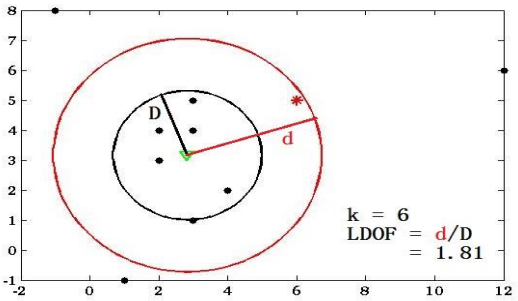

Figure 2: The outlierness of a synthetic two-dimensional i-vector, $\boldsymbol{\omega}_e \in \mathcal{E}$, with respect to its six neighbours, $\boldsymbol{\omega}_p \in \mathcal{P}$, according to the LDOF criteria. Here red star: $\boldsymbol{\omega}_e \in \mathcal{E}$, black dot: $\boldsymbol{\omega}_p \in \mathcal{P}$, green triangle: centre of six $\boldsymbol{\omega}_p \in \mathcal{P}$. Among nine ω_p, three are on the boundary lines.

1. Set the LDOF threshold, θ, so that $0 < \theta \leq 1$.
2. Set $k = 2$.
3. For each $\boldsymbol{\omega}_e \in \mathcal{E}$,
   (a) Find $\mathcal{S}_e^k$.
   (b) Estimate LDOF_e^k.
4. If any $\text{LDOF}_e^k \geq \theta$, then
   (a) $k = k + 1$.
   (b) Go to Step-3
5. Take the unique set of i-vectors from $\{\mathcal{S}_e^k\}_{\forall e}$ to get $\mathcal{S}$.

In order to avoid an extra parameter to tune, we set $\theta = 1$ in our experiments. Fig. 3 shows how the LDOF value is used to decide the value of k in fk-NN for a synthetic two-dimensional enrolment i-vector, $\boldsymbol{\omega}_e$.

4.4. k/fk-NN variants

4.4.1. Individual k-NN (ik-NN)

As mentioned in Subsubsection 4.3.2, it is difficult to estimate an individual k for each $\boldsymbol{\omega}_e$ using LDOF. Here, we propose a variant of fk-NN using the difference between LDOF_e^k's. Let ΔLDOF_e^k be the absolute difference between LDOF_e^k and LDOF_e^{k-1}, and γ be the threshold for ΔLDOF_e^k. Then, for each $\boldsymbol{\omega}_e$, we increase k_e as long as $\text{LDOF}_e^k \geq 1$ and $\Delta\text{LDOF}_e^k \leq \gamma$. Here, we use absolute difference since we assume that the differences converge to 0 as k increases without necessarily being negative for all k. We refer to this method as *individual* k-NN (ik-NN). The steps of ik-NN are given below:

1. For each $\boldsymbol{\omega}_e \in \mathcal{E}$,
   (a) Set $k_e = 2$.
   (b) Find $\mathcal{S}_e^k$.
   (c) Estimate LDOF_e^k.
   (d) If $\text{LDOF}_e^k \geq 1$ and $\Delta\text{LDOF}_e^k \leq \gamma$, then
      i. $k_e = k_e + 1$.
      ii. Go to Step-1b
2. Take the unique set of i-vectors from $\{\mathcal{S}_e^k\}_{\forall e}$ to get $\mathcal{S}$.

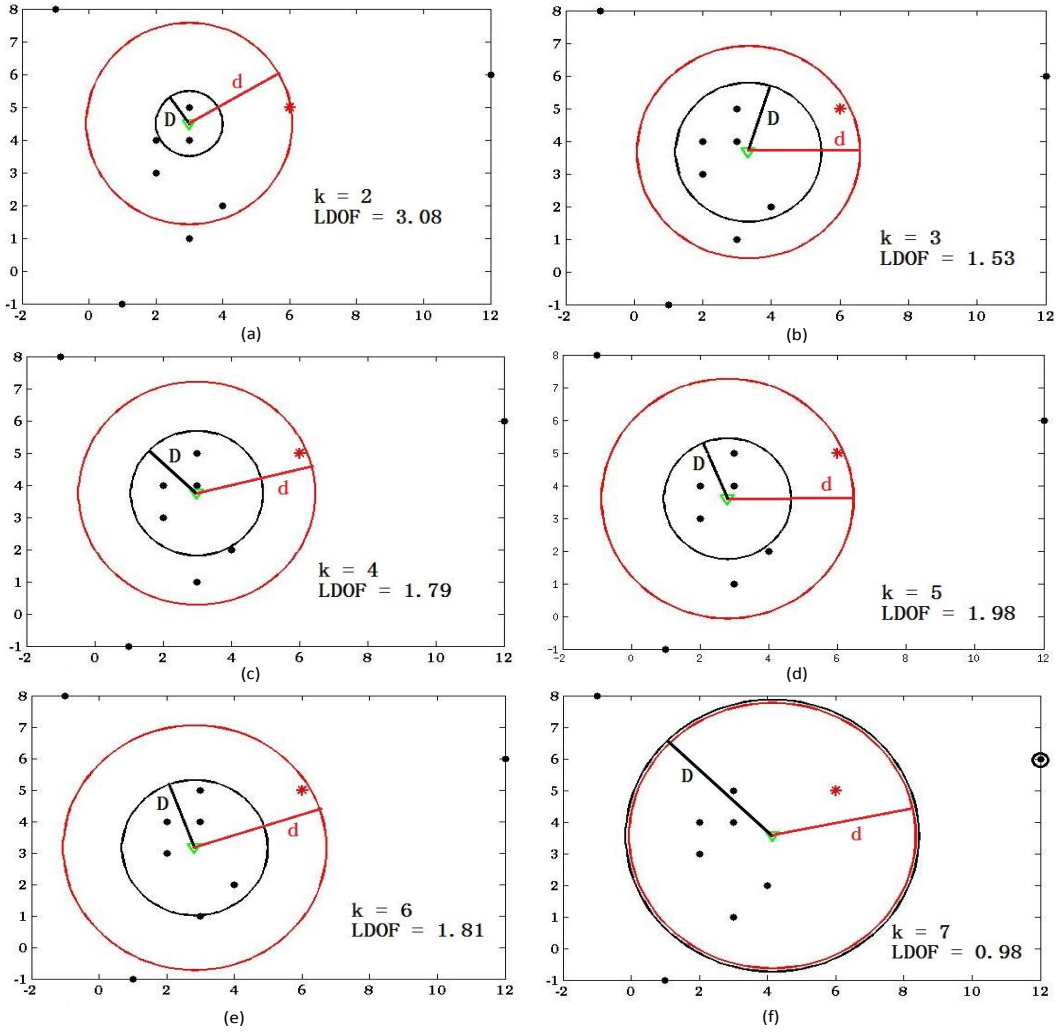

Figure 3: Estimation of k for a synthetic two-dimensional i-vector, $\omega_e \in \mathcal{E}$, by using LDOF value. Here red star: $\omega_e \in \mathcal{E}$, black dot: $\omega_p \in \mathcal{P}$, green triangle: centre of i-vectors in $\mathcal{S}_e^k$.

4.4.2. Averaged enrolment i-vectors

In some cases such as in the NIST SRE12 data set, we sometimes have several enrolment sessions for the same speaker. For such cases, we propose an alternative strategy where at first we average the enrolment i-vectors of each speaker. Then we use the averaged i-vectors for data selection with k-NN or fk-NN in the normal way. We denote the methods as a-k-NN and a-fk-NN, respectively.

4.4.3. Adding all sessions from selected speakers

Having many sessions per speaker is important for reliable estimation of both the speaker variability, $\mathbf{V}$, and the channel variability, $\mathbf{\Sigma}$. Our data selection approaches (k-NN, fk-NN, ik-NN) are, however, unlikely to select all sessions of each selected speaker. We propose a variant of our data selection methods where we first apply k-NN, fk-NN or ik-NN, and then add all discarded i-vectors from the speakers in $\mathcal{S}$. We call this method k-NN-s, fk-NN-s or ik-NN-s. Since adding more i-vectors may lose the theoretical justification for fk-NN-s and ik-NN-s, we focus on k-NN-s in this paper.

4.5. Issues related to data selection

4.5.1. Distance metric

The choice of the distance measure is an important issue in both k-NN and fk-NN. Various measures can be used to compute the distance between two i-vectors. From Fig. 1, we can say that the Euclidean distance could be a good choice. However, since we are using length-normalised i-vectors for PLDA modelling, it would be inconsistent to use the Euclidean distance without length normalisation in the i-vector selection phase. Because, some $\omega_p \in \mathcal{P}$ that are close to an $\omega_e \in \mathcal{E}$ before length-normalisation may not be close after length-normalisation. Thus wrong i-vectors may be selected which will deteriorate the performance of k-NN based system. Our preliminary experiment using the Euclidean distance supported this fact. In Fig. 4, the same i-vectors shown in Fig. 1, are shown after length normalisation. As can be seen, there is an obvious directional separation between the i-vectors extracted from phone- and microphone-utterances. Therefore, the cosine distance could be a good choice. When the i-vectors are length-normalised, the

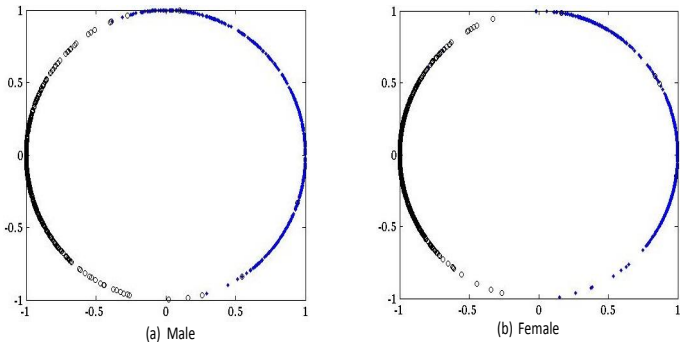

Figure 4: Plot of the length-normalised i-vectors after applying a two dimensional PCA-projection. Black circles represent i-vectors extracted from the phone-utterances and blue dots represent i-vectors extracted from the microphone-utterances. (a) There are 1270 male i-vectors among them 648 are from the phone utterances and 622 are from the microphone utterances. (b) There are 1993 female i-vectors among them 1140 are from the phone-utterances and 853 are from the microphone-utterances.

relation between the two distance metrics is given by,

$$\text{dist}_{\text{euc}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j) = \sqrt{2 \times \text{dist}_{\text{cos}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j)}, \tag{11}$$

where $\text{dist}_{\text{euc}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j)$ is the Euclidean distance and $\text{dist}_{\text{cos}}(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j)$ is the cosine distance between any two i-vectors, $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_j$. Since this function is monotonically rising, it does not make any difference which of the two distance metrics we use when i-vectors are length-normalised. In this study, we use the cosine distance for both k-NN and fk-NN. A more detailed analysis of distance metric will be a part of future work.

4.5.2. Domain adaptation

If we can use $\mathcal{E}$ for selecting relevant data from $\mathcal{P}$, we can also use $\mathcal{E}$ for *domain* adaptation. The most trivial domain adaptation approach is to add i-vectors of $\mathcal{E}$ (i.e., the in-domain data) to the PLDA training data (i.e., the out-domain data), and re-train the model. Domain adaptation adjusts the model to be more similar to $\mathcal{E}$. Data selection improves the modelling in the region close to $\mathcal{E}$ on the expense on regions far from $\mathcal{E}$. Unless the enrolment set is very large, data selection and domain adaptation can be expected to be complementary.

4.5.3. Unseen impostors

A possible concern with the idea of data selection based on $\mathcal{E}$ is that this may reduce the performance for non-target trials involving *unknown* impostors, i.e., impostors who are not in $\mathcal{E}$. The reason for this concern is that by selecting training data close to $\mathcal{E}$, the modelling of impostors in regions far away from $\mathcal{E}$ may deteriorate. However, our assumption is that impostors who are far away from $\mathcal{E}$ will not be confused with speakers in $\mathcal{E}$ anyway. In order to verify that data selection does not reduce the performance for non-target trials involving unknown impostors, we experimentally compare the performance of data selection when all the impostors are unknown and when all the impostors are *known*, i.e., they are one of the speakers in $\mathcal{E}$ in Subsubsection 5.4.6.

5. Experiments

5.1. Outline

We conducted experiments to examine the effect of data set, $\mathcal{S}$, selected by k-NN and fk-NN on the performance of the PLDA model. We restricted our experiments to evaluation sets having an *authentication set*, $\mathcal{A}$, containing telephone data only. We used the NIST SRE 2006 core task (SRE06) as development set, in particular for tuning k in k-NN. We used three evaluation sets, the NIST SRE 2008 core task condition-6 (SRE08), the NIST SRE 2010 core task condition-5 (SRE10), and the NIST SRE 2012 core task condition-2, -4 and -5 (SRE12). Among the three sets, SRE12 has noisy authentication sets, $\mathcal{A}$s, whereas the other two are considered to have clean $\mathcal{A}$s.

During the development of our baseline system, we found that for PLDA training it was beneficial to exclude i-vectors extracted from utterances that were distorted by *echo*, or *crosstalk*, or *background noise* based on meta-data. However, in reality meta-data may not always be available. Therefore, we considered both a training set $\mathcal{R}$, where these *noisy* i-vectors were included, and training set, C, where they were excluded.

The details of our experimental setup are given in Subsection 5.2. The results of SRE06, SRE08 and SRE10 which have clean $\mathcal{A}$s, are given in Subsubsection 5.3.1. The results of SRE12 which has noisy $\mathcal{A}$s are given in Subsubsection 5.3.2. Subsection 5.4 includes analysis of different issues regarding the data selection schemes, as well as experiments with some of their modifications and extensions described in Section 4.

5.2. Experimental set-up

5.2.1. Evaluation sets ($\mathcal{E}$ and $\mathcal{A}$)

The development set, SRE06, and all the evaluation sets, SRE08, SRE10 and SRE12, have an *enrolment set*, $\mathcal{E}$, for training speaker-specific models, an *authentication set*, $\mathcal{A}$, for testing performance of speaker-specific models, and a set of target and non-target trials. Common factors among all evaluation sets are that $\mathcal{A}$ contains data extracted from only conversational telephone speech recorded over ordinary telephone channels and that $\mathcal{E}$ has data extracted from only clean speech files.

SRE06, SRE08 and SRE10 have only clean speech files in $\mathcal{A}$. In $\mathcal{E}$, there are only one speech file for training a model for each target speaker. Each speech file of $\mathcal{E}$ is from approximately five minutes of conversational telephone speech. Some speakers in $\mathcal{E}$ have multiple model IDs. Therefore, the number of speaker models, #M, is larger than the number of speakers, #Es. SRE12 has noisy $\mathcal{A}$s. In SRE12(c5), all speech files of $\mathcal{A}$ have intentionally been collected in a noisy environment. In SRE12(c4), the files of $\mathcal{A}$ have added noise. SRE12(c2) includes all the trials of SRE12(c5) plus trials where $\mathcal{A}$ is clean. In $\mathcal{E}$, multi-session and multi-condition enrolment data are available for each target speaker[3]. There is only one model ID for each speaker in $\mathcal{E}$. Therefore, #M is equal to #Es.

[3]We used the enrolment file list that excludes repeated speech, NIST_SRE12_target_speaker_2_single_file_per_ldcid_map.v2.txt.v2.1.txt.

Table 1: Development set, SRE06 and evaluation sets, SRE08, SRE10 and SRE12 for male and female speakers. #M: the number of models in the enrolment set, $\mathcal{E}$, #Es: the number of unique speakers in $\mathcal{E}$, #Te: the number of test files in the authentication set, $\mathcal{A}$, #As: the number of unique speakers in $\mathcal{A}$, and #Us: the number of speakers of $\mathcal{A}$ unseen in $\mathcal{E}$.

Dataset	Male				
	#M	#Es	#Te	#As	#Us
SRE06	349	257	1347	257	4
SRE08	648	492	858	427	105
SRE10	1906	187	384	192	20
SRE12(c2)	723	723	4962	-	-
SRE12(c4)	723	723	3900	-	-
SRE12(c5)	723	723	2156	-	-
Dataset	Female				
	#M	#Es	#Te	#As	#Us
SRE06	459	335	1679	327	5
SRE08	1140	844	1508	691	92
SRE10	2361	221	369	208	11
SRE12(c2)	1095	1095	7984	-	-
SRE12(c4)	1095	1095	6195	-	-
SRE12(c5)	1095	1095	3325	-	-

Table 1 shows the number of files that we had in $\mathcal{E}$ and $\mathcal{A}$ after discarding corrupted files. In $\mathcal{E}$ of SRE12, there were 8094 and 12393 files for training 723 and 1095 male and female speaker models, respectively. Since the PIN numbers were missing for unknown test segments, #As and #Us for male and female tasks of SRE12 could not be counted. Table 2 shows the number of trials in the evaluation sets. For all sets, only a small number of the non-target trials have impostors who are unseen in $\mathcal{E}$. In more realistic scenarios, the number of unseen impostors might be higher. In SRE12, this is taken into account by a re-balancing of the trials, as will be explained in Subsubsection 5.2.5.

Note that, using $\mathcal{E}$ for system development is allowed in the NIST SRE plan for SRE12 but not for SRE06, SRE08 and SRE10. Therefore, we violated the rules of SRE06, SRE08 and SRE10.

5.2.2. Training data of UBM and T

For training UBM and **T** matrix, we used the NIST SRE 2004 (SRE04), NIST SRE 2005 (SRE05), Switchboard II Phase 1 (SB2P1), Switchboard II Phase 2 (SB2P2), Switchboard II Phase 3 (SB2P3), Switchboard Cellular Part 1 (SBCP1) and Switchboard Cellular Part 2 (SBCP2). From SRE04, we selected speech files having single-channel conversation of approximately five minutes total duration. From SRE05, we selected speech files having two-channel conversation of approximately five minutes total duration. We used all non-empty speech files of the Switchboard datasets.

The number of speech files, #F, and the number of speakers, #S, used for training UBMs and **T** matrices are shown in Table 3. MIXER PIN and PIN were used as unique speaker IDs for NIST SRE and Switchboard datasets respectively. For the files whose MIXER PIN or PIN were missing, model IDs were used as speaker IDs. For example, in SRE05, there were 198 male and 211 female speech segments without MIXER PIN. We counted those speech segments as from 28 male and 29 female speakers based on their model IDs. However, multiple model IDs may share the same MIXER PIN. Therefore, it is possible that our counted #S was higher than the original #S. There were 18 male and 23 female speakers appearing in multiple Switchboard datasets. Therefore, the number of speakers in the combined set, $\mathcal{R}$, was smaller than the total speakers of individual sets. For male set, #S of $\mathcal{R}$ was 1495, whereas total #S was 1525. For female set, #S of $\mathcal{R}$ was 1897, whereas total #S was 1919.

Table 2: Trials of SRE06, SRE08, SRE10 and SRE12 for male and female speakers. #T: the number of total trials, #Tr: the number of target trials, #Nt: the number of non-target trials, #Kn: the number of non-target trials by known speakers, #Un: the number of non-target trials by unknown speakers.

Dataset	Male				
	#T	#Tr	#Nt	#Kn	#Un
SRE06	22123	1594	20529	20066	463
SRE08	12356	724	11632	9906	1726
SRE10	179338	3465	175873	158846	17027
SRE12(c2)	164549	2830	161719	131932	29787
SRE12(c4)	125400	2775	122625	122625	0
SRE12(c5)	62845	1534	61311	61311	0
Dataset	Female				
	#T	#Tr	#Nt	#Kn	#Un
SRE06	28945	2022	26923	26478	445
SRE08	22957	1445	21512	20088	1424
SRE10	236781	3704	233077	221097	11980
SRE12(c2)	393042	4524	388518	313109	75409
SRE12(c4)	298491	4401	294090	289218	4872
SRE12(c5)	152976	2349	150627	148221	2406

Table 3: The number of speech files, #F, and the number of speakers, #S, used for training gender-dependent UBM and **T**.

Dataset	Male		Female	
	#F	#S	#F	#S
SB2P1	2558	292	3251	358
SB2P2	2352	304	2716	335
SB2P3	1612	290	2083	341
SBCP1	462	103	567	116
SBCP2	1310	165	2000	245
SRE04	1906	126	2651	188
SRE05	2705	245	3792	336
$\mathcal{R}$	12905	1495	17060	1897

5.2.3. Training datasets for PLDA model ($\mathcal{R}$ & C)

For training PLDA models, we prepared two sets of speech files, $\mathcal{R}$ and C. In $\mathcal{R}$, we included all speech files of $\mathcal{P}$, i.e., the files used for training UBM and **T** matrices. For C, we selected only the *clean speech* files of $\mathcal{P}$. Clean speech refers to speech which is not distorted by *echo* or *crosstalk* or *back-*

Table 4: The number of speech files, #F, and the number of speakers, #S, selected from clean speech for training gender-dependent PLDA models.

Dataset	Male		Female	
	#F	#S	#F	#S
SB2P1	391	125	455	191
SB2P2	1868	283	2134	307
SB2P3	1399	277	1921	337
SBCP1	236	78	290	94
SBCP2	1038	157	1595	232
SRE04	1906	126	2651	188
SRE05	2705	245	3792	336
C	9543	1278	12838	1665

ground noise according to the meta-data of the databases. According to the documentation of the Switchboard corpora (Graff et al., 1998), echo or crosstalk in the telephone circuit refers to the audibility of the channel-1 speaker in channel-2 and vice-versa. Background noise refers to the amount of sounds not made by the speakers, e.g., baby crying, television, radio, etc. For the NIST SRE databases, there is no meta-data for identifying *noisy* speech, therefore we considered all speech of SRE04 and SRE05 as clean speech. The number of speakers, #S, and the number of clean speech files, #F, of individual dataset and of the combined dataset, C, are given in Table 4.

5.2.4. Pre-processing and training models

We at first extracted 15 PLP coefficients (Hermansky, 1990) along with log-energy and then applied feature warping (Pelecanos & Sridharan, 2001). After that we appended the first-order and second-order derivatives, resulting in 48 elements per frame. Then we removed non-speech parts from the feature vector sequences by using spectral subtraction-based voice activity detector (VAD) (Mak & Yu, 2010).

After extracting PLP features, we trained gender-dependent systems. First, we trained gender-dependent UBMs with 2048 Gaussian components by using feature vectors of $\mathcal{R}$. Then we estimated sufficient statistics. Next we trained gender-dependent $\mathbf{T}$ matrices by using sufficient statistics estimated by the feature vectors extracted from $\mathcal{R}$. The rank of $\mathbf{T}$ matrices, d, was tuned to 400 by using SRE06. By using $\mathbf{T}$ matrices, we extracted i-vectors of $\mathcal{R}$. Then, we applied data selection and domain adaptation for selecting i-vectors for training PLDA models. Finally, the i-vectors of PLDA models went through the process of centring, whitening, and length-normalisation (Garcia-Romero & Espy-Wilson, 2011).

We trained gender-dependent PLDA models. The parameters $\mathbf{m}$, $\mathbf{V}$ and $\mathbf{\Sigma}$ of PLDA models were estimated by the ML criteria. The rank of $\mathbf{V}$ was optimised to 250 by using SRE06. Table 5 defines the symbols for referring to the data sets we used in the experiments. Note that we will use the same symbol for the training set and its corresponding PLDA model from now on.

Table 5: Symbols that will be used for referring to PLDA models later in this paper.

Symbol	Training Data
$\mathcal{R}$	All available data
C	Clean data selected by removing echo or crosstalk or noise from $\mathcal{R}$, i.e., $\{C \subset R\}$
$\{C/\mathcal{R}\}_{\mathrm{k}}$	$\mathcal{S} \subset \{C/\mathcal{R}\}$ selected by k-NN
$\{C/\mathcal{R}\}_{\mathrm{fk}}$	$\mathcal{S} \subset \{C/\mathcal{R}\}$ selected by fk-NN
C_{ik}	$\mathcal{S} \subset C$ selected by ik-NN
$\{C/\mathcal{R}\} + \mathcal{E}$	Training data added with enrolment set
$\{C/\mathcal{R}\}_{\mathrm{k}} + \mathcal{E}$	$\mathcal{S} \subset \{C/\mathcal{R}\}$ selected by k-NN and added with $\mathcal{E}$
$\{C/\mathcal{R}\}_{\mathrm{fk}} + \mathcal{E}$	$\mathcal{S} \subset \{C/\mathcal{R}\}$ selected by fk-NN and added with $\mathcal{E}$

5.2.5. Performance measure

For SRE06, SRE08 and SRE10, we used *equal error rate* (EER) and *minimum detection cost*, $\mathrm{C}^{\mathrm{min}}$, as evaluation metrics. The EER indicates the number of errors when the decision threshold is set so that the proportion of false acceptances (FA) and the proportion of false rejections (FR) are equal. The $\mathrm{C}^{\mathrm{min}}$ is the minimum value of normalised *detection cost function*, $\mathrm{C}_{\mathrm{Norm}}$, defined as:

$$\mathrm{C}_{\mathrm{Norm}} = \mathrm{C}_{\mathrm{Det}}/\mathrm{C}_{\mathrm{Default}}, \tag{12}$$

where $\mathrm{C}_{\mathrm{Det}}$ is a weighted sum of FR and FA error probabilities, and $\mathrm{C}_{\mathrm{Default}}$ is the best cost that could be obtained either by always accepting or always rejecting the segment speaker as matching the target speaker, whichever gives the lower cost. According to the NIST SRE plans for SRE06, SRE08 and SRE10 (NIST, 2006, 2008, 2010), $\mathrm{C}_{\mathrm{Default}}$ and $\mathrm{C}_{\mathrm{Det}}$ can be defined as :

$$\mathrm{C}_{\mathrm{Default}} = \min \left\{ \begin{matrix} \mathrm{C}_{\mathrm{FR}} \times \mathrm{P}_{\mathrm{Target}}, \\ \mathrm{C}_{\mathrm{FA}} \times (1 - \mathrm{P}_{\mathrm{Target}}) \end{matrix} \right\}, \tag{13}$$

and

$$\begin{aligned} \mathrm{C}_{\mathrm{Det}} =& \mathrm{C}_{\mathrm{FR}} \times \mathrm{P}_{\mathrm{FR|Target}} \times \mathrm{P}_{\mathrm{Target}} \\ &+ \mathrm{C}_{\mathrm{FA}} \times \mathrm{P}_{\mathrm{FA|Nontarget}} \times (1 - \mathrm{P}_{\mathrm{Target}}), \end{aligned} \tag{14}$$

where C_{FR} and C_{FA} are the relative costs of detection errors, and $\mathrm{P}_{\mathrm{Target}}$ is the *a priori* probability of the specified target speaker. For SRE06 and SRE08, $\mathrm{C}_{\mathrm{FR}} = 10, \mathrm{C}_{\mathrm{FA}} = 1$ and $\mathrm{P}_{\mathrm{Target}} = 0.01$. On the other hand, for SRE10, $\mathrm{C}_{\mathrm{FR}} = \mathrm{C}_{\mathrm{FA}} = 1$ and $\mathrm{P}_{\mathrm{Target}} = 0.001$.

For SRE12, we used an actual and a minimum version of the primary evaluation metric, denoted by $\mathrm{C}^{\mathrm{act}}$ and $\mathrm{C}^{\mathrm{min}}$, respectively, as evaluation metrics. The primary evaluation metric of SRE12, $\mathrm{C}_{\mathrm{Primary}}$, is the average of two normalised detection costs, $\mathrm{C}^{(1)}_{\mathrm{Norm}}$ and $\mathrm{C}^{(2)}_{\mathrm{Norm}}$, given by

$$\mathrm{C}_{\mathrm{Primary}} = \frac{\mathrm{C}^{(1)}_{\mathrm{Norm}} + \mathrm{C}^{(2)}_{\mathrm{Norm}}}{2}, \tag{15}$$

where

$$C^{(i)}_{\text{Norm}} = P_{\text{FR|Target}} + \beta^{(i)} \times P_{\text{Known}} \times P_{\text{FA|KnownNontarget}} + \beta^{(i)} \times (1 - P_{\text{Known}}) \times P_{\text{FA|UnknownNontarget}}, \quad (16)$$

where P_{Known} is the *a priori* probability that the non-target speaker is one of the enrolled speakers, and

$$\beta^{(i)} = \frac{C_{\text{FR}}}{C_{\text{FA}}} \times \frac{P^{(i)}_{\text{Target}}}{1 - P^{(i)}_{\text{Target}}}; \text{ for i=1,2.} \quad (17)$$

For both $C^{(1)}_{\text{Norm}}$ and $C^{(2)}_{\text{Norm}}$, $C_{\text{FR}} = C_{\text{FA}} = 1$. The prior probability for a target trial is $P^{(1)}_{\text{Target}} = 0.01$ and $P^{(2)}_{\text{Target}} = 0.001$ for $C^{(1)}_{\text{Norm}}$ and $C^{(2)}_{\text{Norm}}$, respectively.

We computed C^{act} by applying detection thresholds of $\log(\beta)$ for the two values of β with $\beta^{(1)} = 99$ and $\beta^{(2)} = 999$ as recommended in NIST SRE plan for SRE12. C^{min} was the C_{Primary} estimating by averaging the minimum versions of $C^{(1)}_{\text{Primary}}$ and $C^{(2)}_{\text{Primary}}$. When $P_{\text{Known}} > 0$, we used compound LLRs instead of the original *simple* LLRs.[4] In order for compound LLRs to be effective, it is important that the simple LLRs are well-calibrated. It is not sufficient to calibrate the compound LLRs themselves. Therefore, to simply optimise the decision threshold for the compound LLRs does not give the lowest cost that could have been obtained with perfect calibration. For C^{min}, we, therefore, trained and applied a PAV transformation on the evaluation scores. For C^{act}, we used an affine transformation estimated using the C_{llr} loss shifted to $P_{\text{Target}} = 10^{-2.5}$ (i.e., the geometric average of $P^{(1)}_{\text{Target}}$ and $P^{(2)}_{\text{Target}}$). We used SRE06 for training the affine transformation. For calculating compound LLRs, doing calibration and calculating the evaluation metrics, we used the BOSARIS toolbox (Brümmer, 2012).

5.2.6. Tuning k

For the conventional k-NN, we optimised k by minimising EER of the development set, SRE06. Using the cosine distance as the distance metric, we chose the k-nearest neighbours from C for each $\omega_e \in \mathcal{E}$. We increased k from one up to fifty. When $k \leq 2$, PLDA training failed due to an insufficient amount of training data. The optimum k was 37 for male and 25 for female trials of SRE06, respectively. We used the same k for the training set, $\mathcal{R}$.

5.3. Results

5.3.1. SRE06, SRE08 and SRE10

Table 6 compares EER and C^{min} for the baseline, k-NN and fk-NN. Data selection either by k-NN or by fk-NN improved the verification accuracy. The k-NN method performed well on the development set, SRE06, where k was optimised. On the other hand, fk-NN was better than k-NN for reducing EER in SRE08 and SRE10. Using fk-NN in C, we achieved on average 4.2% and 3.4% relative reduction in EER for male and female trials of the evaluation sets, SRE08 and SRE10, respectively. Using fk-NN in $\mathcal{R}$, we achieved on average 6.0% and 6.6% relative reduction in EER for male and female trials of the evaluation sets, SRE08 and SRE10, respectively. Compared to the clean set, C, we were more successful in reducing EER by selecting i-vectors from $\mathcal{R}$ using fk-NN.

Table 6: EER and C^{min} of SRE06, SRE08 and SRE10. For SRE06 and SRE08, C^{min} is in 10^{-2} whereas for SRE10, C^{min} is in 10^{-4}. For all tasks EER is in %.

Male model	SRE06		SRE08		SRE10	
	EER	C^{min}	EER	C^{min}	EER	C^{min}
C	2.30	1.16	4.92	2.55	2.01	3.73
C_{k}	**1.84**	**1.05**	4.76	2.44	2.05	3.68
C_{fk}	2.08	1.12	**4.73**	**2.43**	**1.92**	**3.53**
$\mathcal{R}$	2.59	1.33	5.07	2.65	2.14	3.97
$\mathcal{R}_{\text{k}}$	2.08	**1.13**	4.87	**2.58**	2.11	3.94
$\mathcal{R}_{\text{fk}}$	**2.07**	1.15	**4.77**	**2.58**	**2.01**	**3.76**
Female model	SRE06		SRE08		SRE10	
	EER	C^{min}	EER	C^{min}	EER	C^{min}
C	3.42	1.85	5.97	2.85	3.02	4.94
C_{k}	**2.71**	**1.43**	5.81	**2.82**	2.93	**4.74**
C_{fk}	**2.71**	**1.43**	**5.78**	2.84	**2.91**	**4.74**
$\mathcal{R}$	3.92	2.20	6.29	3.01	3.29	4.96
$\mathcal{R}_{\text{k}}$	**2.89**	**1.50**	5.80	**2.84**	3.22	4.83
$\mathcal{R}_{\text{fk}}$	**2.89**	**1.50**	**5.79**	**2.84**	**3.16**	**4.81**

5.3.2. SRE12

SRE12 is different from the previous evaluation sets in that several enrolment sessions are available for each speaker in $\mathcal{E}$. In our preliminary experiments, the methods using the averaged i-vector methods, a-k-NN and a-fk-NN (4.4.2), were always better than the methods using all available i-vectors per speaker. Therefore, we considered only a-k-NN and a-fk-NN in these experiments.

The results of data selection from C and $\mathcal{R}$ are given in Table 7 and Table 8. Since SRE12(c4) and SRE12(c5) for male do not have any unknown impostors, the performance for these conditions could not be estimated. With the exception of using C for male, data selection was always effective. In many cases, k-NN was better than fk-NN. A possible reason for this could be that averaged i-vectors were not optimal for determining k with fk-NN. It is noticeable that, despite the fact that $\mathcal{A}$ was noisy, using C gave in most cases better C^{act} than using $\mathcal{R}$. This could perhaps be explained by the fact that the calibration model was trained on SRE06 which was clean.

5.4. Analysis

In this subsection we analyse the behaviour of data selection with k-NN and fk-NN more in detail, as well as some of their modifications and extensions described in Section 4. Most of the experiments were done on SRE06, SRE08 and SRE10. Only for checking the effect of unseen impostors, SRE12 was used.

[4] See https://sites.google.com/site/bosaristoolkit/sre12 and the materials therein for an explanation about compound vs. simple LLRs.

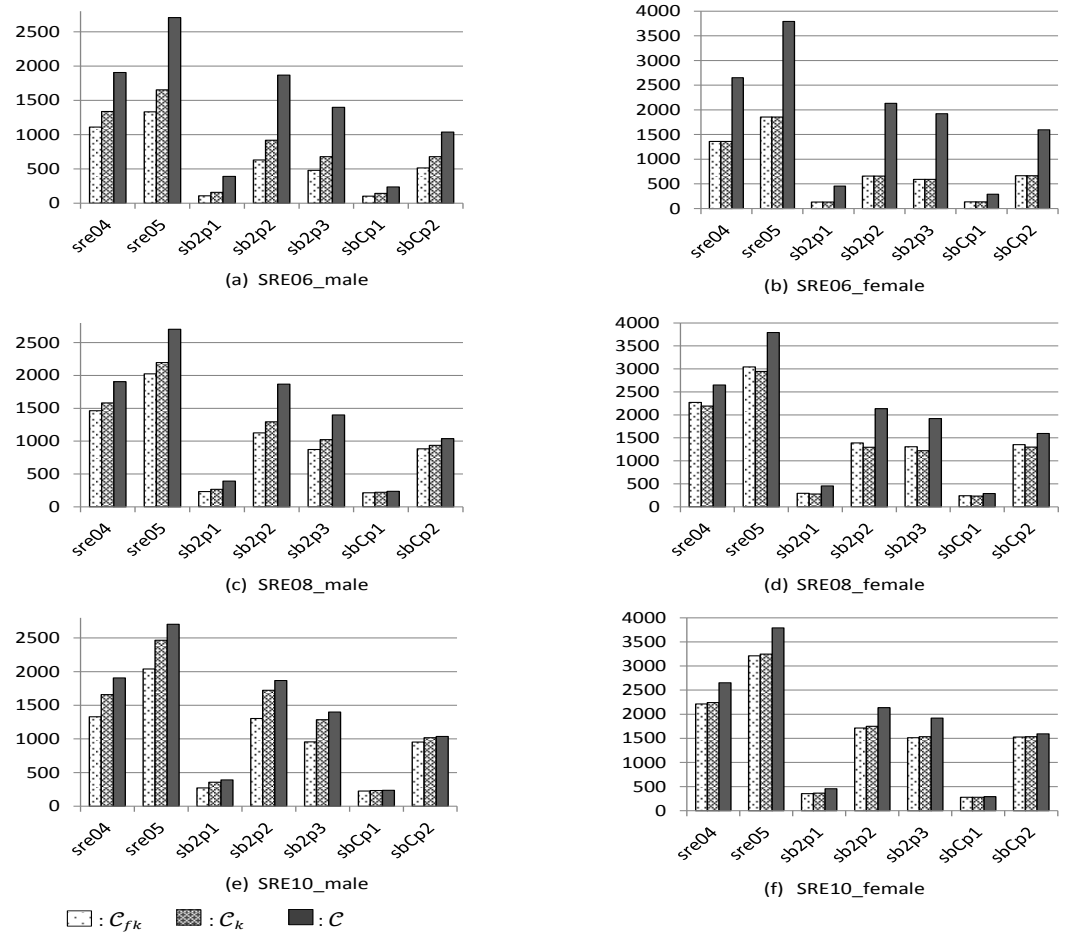

Figure 5: The *y*-axis shows the number of i-vectors of different datasets used for training PLDA models for male and female trials of SRE06, SRE08 and SRE10.

Table 7: Results of male trials of SRE12(c2) using $P_{Known} = 0.5$. In *k*-NN, *k* was optimised considering SRE06 as the development set.

PLDA model	C^{act}	C^{min}
$\mathcal{C}$	0.321	**0.293**
$\mathcal{C}_{\text{a-k}}$	**0.320**	0.298
$\mathcal{C}_{\text{a-fk}}$	0.325	0.315
$\mathcal{R}$	0.350	0.287
$\mathcal{R}_{\text{a-k}}$	0.331	0.289
$\mathcal{R}_{\text{a-fk}}$	**0.323**	**0.279**

Table 8: Results of female trials of SRE12 using $P_{Known} = 0.5$. In *k*-NN, *k* was optimised considering SRE06 as the development set.

PLDA	c2		c4		c5	
model	C^{act}	C^{min}	C^{act}	C^{min}	C^{act}	C^{min}
$\mathcal{C}$	0.432	0.281	0.598	0.464	0.491	0.310
$\mathcal{C}_{\text{a-k}}$	**0.386**	**0.271**	**0.548**	**0.447**	**0.436**	0.308
$\mathcal{C}_{\text{a-fk}}$	0.414	0.279	0.574	0.452	0.470	**0.307**
$\mathcal{R}$	0.476	0.281	0.630	0.444	0.536	0.317
$\mathcal{R}_{\text{a-k}}$	**0.409**	0.277	**0.558**	**0.441**	**0.461**	0.314
$\mathcal{R}_{\text{a-fk}}$	0.445	**0.273**	0.596	0.445	0.504	**0.299**

5.4.1. Analysis of the selected data

Fig. 5(a)-5(f) shows how much data were selected from each training corpus for SRE06, SRE08 and SRE10. The most noticeable trend was that SRE10 selected much more of the Switchboard corpora than SRE06 and SRE08. In particular, SRE10 selected almost all of SBCP1. For further analysis, we trained database-specific PLDA models using data of each database in $\mathcal{C}$. For SB2P1 and SBCP1, PLDA training failed due to an insufficient amount of training data.[5] The results are shown in Table 9. We can conclude that the NIST SRE databases (ALLSRE) have more relevant data for $\mathcal{E}$ of SRE06, SRE08, and SRE10 than the Switchboard databases (ALLSB). Using only ALLSRE, we got the lowest EER and C^{min} for SRE06 and SRE08 while adding ALLSB with ALLSRE had negative impact on the performance. It reveals that using all the available data does not guarantee the best PLDA model for the target evaluation set. The presence of irrelevant data in the training set of the PLDA model may deteriorate the system's performance. For SRE10, we got the lowest EER and C^{min} when we combined ALLSB with ALLSRE. It indicates that relevant data differs in different target evaluation sets. By *k*-NN and fk-NN, we

[5] In this study, we did not attempt to solve this problem by applying regularisation to the channel covariance during PLDA training.

Table 9: EER and C^{min} of SRE06, SRE08 and SRE10 for different training data sets. Empty entries mean that PLDA training failed due to insufficient amount of training data. For SRE06 and SRE08, C^{min} is in 10^{-2} whereas for SRE10, C^{min} is in 10^{-4}. For all tasks, EER is in %.

Male model	SRE06		SRE08		SRE10	
	EER	C^{min}	EER	C^{min}	EER	C^{min}
SB2P1	-	-	-	-	-	-
SB2P2	10.65	4.99	13.26	5.54	17.96	8.56
SB2P3	11.23	5.28	14.31	5.65	18.19	8.78
SBCP1	-	-	-	-	-	-
SBCP2	16.88	6.0	15.16	5.79	12.47	9.81
ALLSB	8.17	3.86	9.65	4.83	5.38	6.54
SRE04	4.89	2.20	6.96	3.29	4.88	7.51
SRE05	3.85	1.94	5.74	2.99	2.81	5.65
ALLSRE	**2.07**	**0.97**	**4.58**	**2.36**	2.28	4.20
C	2.30	1.16	4.92	2.55	**2.01**	**3.73**
Female model	SRE06		SRE08		SRE10	
	EER	C^{min}	EER	C^{min}	EER	C^{min}
SB2P1	-	-	-	-	-	-
SB2P2	12.56	6.0	14.94	6.42	17.92	9.15
SB2P3	11.57	6.07	13.83	6.24	17.42	8.77
SBCP1	-	-	-	-	-	-
SBCP2	11.20	5.34	11.43	5.32	7.56	8.54
ALLSB	9.04	5.16	11.12	5.41	5.88	6.92
SRE04	3.35	1.73	6.53	3.00	4.62	6.83
SRE05	5.06	2.60	6.94	3.25	3.74	5.26
ALLSRE	**2.64**	**1.42**	**5.51**	**2.6**	3.26	**4.69**
C	3.42	1.85	5.97	2.85	**3.02**	4.94

are able to reduce the amount of irrelevant data for the target evaluation set.

5.4.2. Error analysis

In order to analyse the errors, we counted the number of *false acceptance* (FA) and *false rejection* (FR) as well as the number of enrolment and test segments that had at least one erroneous decision for any trial in SRE06, SRE08 and SRE10. For this analysis, we used the thresholds that minimised the detection costs. For the baseline system, the number of FR was higher than the number of FA. This is because the operating point of C^{min} in SRE06, SRE08 and SRE10 promotes a low FA rate. This is particularly extreme for SRE10. As shown in Table 10, we noticed that data selection reduced the number of FR, miss-recognised speakers and test-segments in all datasets except SRE08. In most of the cases, the number of FA increased when data selection was applied.

5.4.3. Adding all sessions from selected speakers

According to the speaker based data selection approach, k-NN-s, described in Subsubsection 4.4.3, all the sessions of each selected speaker were added in to $\mathcal{S}$ and k was optimised. The optimal value of k on SRE06 was 12 and 3 for male and female tasks, respectively, compared to 37 and 25 for the standard approach. Table 11 shows the results. As can be seen, this method

Table 10: Number of errors. FR: False Rejection, FA : False Acceptance, eS: Erroneous Target Speakers, eT: Erroneous Test Segments.

SRE06	Male			Female		
	C	C_k	C_{fk}	C	C_k	C_{fk}
FR	137	113	114	244	209	209
FA	84	91	105	219	142	142
eS	124	112	117	226	194	194
eT	194	178	186	380	303	303
SRE08	Male			Female		
	C	C_k	C_{fk}	C	C_k	C_{fk}
FR	137	92	96	274	266	269
FA	104	163	155	265	269	267
eS	171	184	179	345	345	350
eT	196	194	186	405	405	404
SRE10	Male			Female		
	C	C_k	C_{fk}	C	C_k	C_{fk}
FR	1156	1098	1046	1529	1518	1438
FA	7	9	9	19	15	20
eS	798	768	737	1125	1118	1075
eT	253	252	246	290	281	276

Table 11: EER and C^{min} for speaker based i-vector selection. k was tuned on SRE06. The "*" indicates that k was optimised for this method. In the other rows, k was optimised before adding discarded sessions of the selected speakers. For SRE06 and SRE08, C^{min} is in 10^{-2} whereas for SRE10, C^{min} is in 10^{-4}. For all tasks, EER is in %.

Male model	SRE06		SRE08		SRE10	
	EER	C^{min}	EER	C^{min}	EER	C^{min}
C	2.30	1.16	4.92	2.55	**2.01**	**3.73**
$C_{k\text{-s}}, k = 37$	2.28	1.21	5.01	2.58	2.08	3.93
$C^*_{k\text{-s}}, k = 12$	**2.03**	**1.12**	**4.86**	**2.48**	2.04	3.77
$C_{fk\text{-s}}$	2.22	1.19	4.89	2.54	2.08	3.88
Female model	SRE06		SRE08		SRE10	
	EER	C^{min}	EER	C^{min}	EER	C^{min}
C	3.42	1.85	5.97	2.85	**3.02**	4.94
$C_{k\text{-s}}, k = 25$	3.50	1.96	6.20	2.98	3.19	4.95
$C^*_{k\text{-s}}, k = 3$	**2.71**	**1.43**	**5.79**	**2.82**	3.05	**4.75**
$C_{fk\text{-s}}$	3.50	1.96	6.17	2.99	3.20	4.90

did not perform well with the values of k that were optimal for the standard approach. However, when k was specifically optimised for this purpose, the result was comparable to the standard approach. This approach could, however, be refined by ensuring that every speaker has at least a certain number of sessions rather than using all the available sessions. Such exploration will be a part of future work.

5.4.4. Domain adaptation

Table 12 compares the baseline, k-NN and fk-NN when $\mathcal{E}$ was added to the PLDA training set. Notice that $\mathcal{E}$ was added after data selection. The addition of $\mathcal{E}$ improved the performance of all systems substantially. For male trials of SRE06, SRE08 and SRE10, by using $C + \mathcal{E}$ we achieved 28.7%, 15.9%

Table 12: EER and C^{min} for systems trained by including $\mathcal{E}$ into $\mathcal{P}$. In k-NN, k was optimised using SRE06. For male, $k = 37$, and for female, $k = 25$. For SRE06 and SRE08, C^{min} is in 10^{-2} whereas for SRE10, C^{min} is in 10^{-4}. For all tasks EER is in %.

Male model	SRE06		SRE08		SRE10	
	EER	C^{min}	EER	C^{min}	EER	C^{min}
$C + \mathcal{E}$	1.64	0.90	4.14	2.00	1.46	2.97
$C_k + \mathcal{E}$	**1.35**	**0.76**	**3.92**	1.92	1.48	3.05
$C_{fk} + \mathcal{E}$	1.53	0.77	3.97	**1.80**	**1.46**	**2.91**
$\mathcal{R} + \mathcal{E}$	1.88	1.01	4.41	2.27	1.61	3.21
$\mathcal{R}_k + \mathcal{E}$	**1.52**	**0.81**	4.04	**2.06**	**1.55**	3.26
$\mathcal{R}_{fk} + \mathcal{E}$	1.54	**0.81**	**3.96**	2.08	**1.55**	**2.95**
Female model	SRE06		SRE08		SRE10	
	EER	C^{min}	EER	C^{min}	EER	C^{min}
$C + \mathcal{E}$	2.41	1.29	5.09	2.21	2.51	4.31
$C_k + \mathcal{E}$	**2.09**	**1.01**	**4.74**	2.12	2.36	4.06
$C_{fk} + \mathcal{E}$	**2.09**	**1.01**	4.85	**2.08**	**2.34**	**4.04**
$\mathcal{R} + \mathcal{E}$	2.90	1.47	5.49	2.42	2.68	4.42
$\mathcal{R}_k + \mathcal{E}$	**2.21**	**1.03**	**4.96**	**2.21**	2.57	**4.14**
$\mathcal{R}_{fk} + \mathcal{E}$	**2.21**	**1.03**	4.99	2.23	**2.55**	4.16

and 27.4% relative reduction in EER over C, respectively. For female trials of SRE06, SRE08 and SRE10, the EER reduction rates were 29.5%, 14.7% and 16.9%, respectively. These results confirmed the effect of domain adaptation.

We observed a consistent improvement using data-selection followed by domain adaptation. By using $C_{fk} + \mathcal{E}$, we achieved 6.7% and 4.1% EER reduction over $C + \mathcal{E}$ for male trials of SRE06 and SRE08, respectively. For female trials of SRE06, SRE08 and SRE10, the EER reduction rates were 13.3%, 4.7% and 6.8%, respectively. Fig. 6 shows the DET curves. It is clear that adding $\mathcal{E}$ to C improved PLDA modelling and that fk-NN improved the system performance further by discarding irrelevant data from $\mathcal{P}$.

When $\mathcal{E}$ was included with $\mathcal{R}$, using fk-NN, the EER reduction rates were 18.1%, 10.2% and 3.7%, respectively, for male trials of SRE06, SRE08 and SRE10. For female trials of SRE06, SRE08 and SRE10, the EER reduction rates were 23.8%, 9.1% and 4.9%, respectively. We can conclude that $\{\mathcal{P}_{fk} + \mathcal{E}$ or $\mathcal{P}_k + \mathcal{E}\} > \{\mathcal{P} + \mathcal{E}\} >> \{\mathcal{P}_{fk}$ or $\mathcal{P}_k\} > \{\mathcal{P}\}$, where $>$ refers *better* and $>>$ refers *much better* performance.

5.4.5. Individual k-NN

In all of our experiments up until now, we used the same k for all $\omega_e \in \mathcal{E}$. Here, we show experiment with *ik*-NN proposed in Subsubsection 4.4.1. Table 13 shows results of using $\gamma = 0.0001$. Overall, *ik*-NN outperformed our baselines systems, but it was not better than fk-NN. A comparison of this method and the standard k-NN for different amounts of training data is shown in Figure 7. For k-NN, the amount of training data was controlled by varying the value of k, and for *ik*-NN the amount of training data was controlled by varying the threshold, γ. For smaller training data sizes, ik-NN was better but for larger sizes, k-NN was better. Both the methods reached, however, a similar optimum. Using ω_e dependent k is tricky and the proposed *ik*-NN is unlikely to be optimal. Exploring other strategies may, therefore, be a fruitful direction of future work.

Table 13: EER and C^{min} for systems trained by C, and C_{ik}. For both male and female, $\gamma = 0.0001$. For SRE06 and SRE08, C^{min} is in 10^{-2} whereas for SRE10, C^{min} is in 10^{-4}. For all tasks EER is in %.

Male model	SRE06		SRE08		SRE10	
	EER	C^{min}	EER	C^{min}	EER	C^{min}
C	2.30	1.16	4.92	2.55	2.01	3.73
C_k	**1.84**	**1.05**	4.76	**2.44**	2.05	3.68
C_{fk}	2.08	1.12	4.73	2.43	**1.92**	**3.53**
C_{ik}	1.86	1.12	**4.54**	2.46	2.00	3.72
Female model	SRE06		SRE08		SRE10	
	EER	C^{min}	EER	C^{min}	EER	C^{min}
C	3.42	1.85	5.97	2.85	3.02	4.94
C_k	**2.71**	**1.43**	5.81	**2.82**	2.93	**4.74**
C_{fk}	**2.71**	**1.43**	**5.78**	2.84	**2.91**	**4.74**
C_{ik}	2.93	1.46	5.83	2.84	2.93	4.77

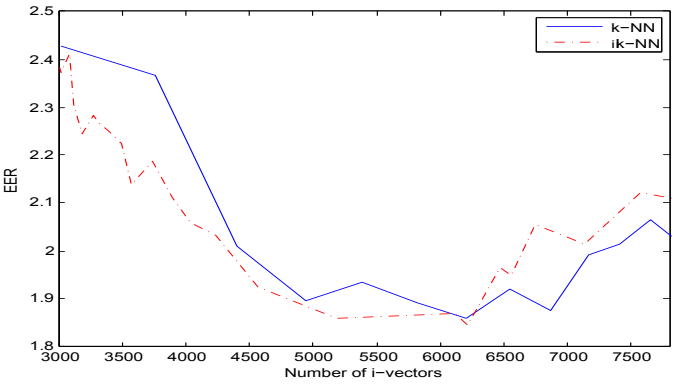

Figure 7: EER(%) of SRE06, male. The x-axis shows the number of i-vectors selected by k-NN and *ik*-NN for training the PLDA model.

5.4.6. Effect on unseen impostors

As discussed in Subsubsection 4.5.3, we need to confirm whether data selection has a bad effect on unknown impostors. For this, we examined the performance of k-NN and fk-NN on SRE12(c2) when $P_{Known} = 1$ and $P_{Known} = 0$. The results are given in Table 14. Overall, the performance of all methods became better when $P_{Known} = 1$, since we used compound LLRs that took advantage of the presence of known impostors. When $P_{Known} = 0$, data selection resulted in large improvements for female but a less clear pattern for male. However, notice in Table 2 that the number of trials from unknown impostors is quite small for male, so these results might be less reliable. In conclusion, it seems unknown non-target trials are not problematic for our data selection schemes.

5.4.7. Data reduction rate

Table 15 shows the data reduction rates for the four data sets. It is clear that more irrelevant data was reduced from $\mathcal{R}$ than C by both k-NN and fk-NN. For the male sets, fk-NN reduced the data more than k-NN. For SRE10, both k-NN and

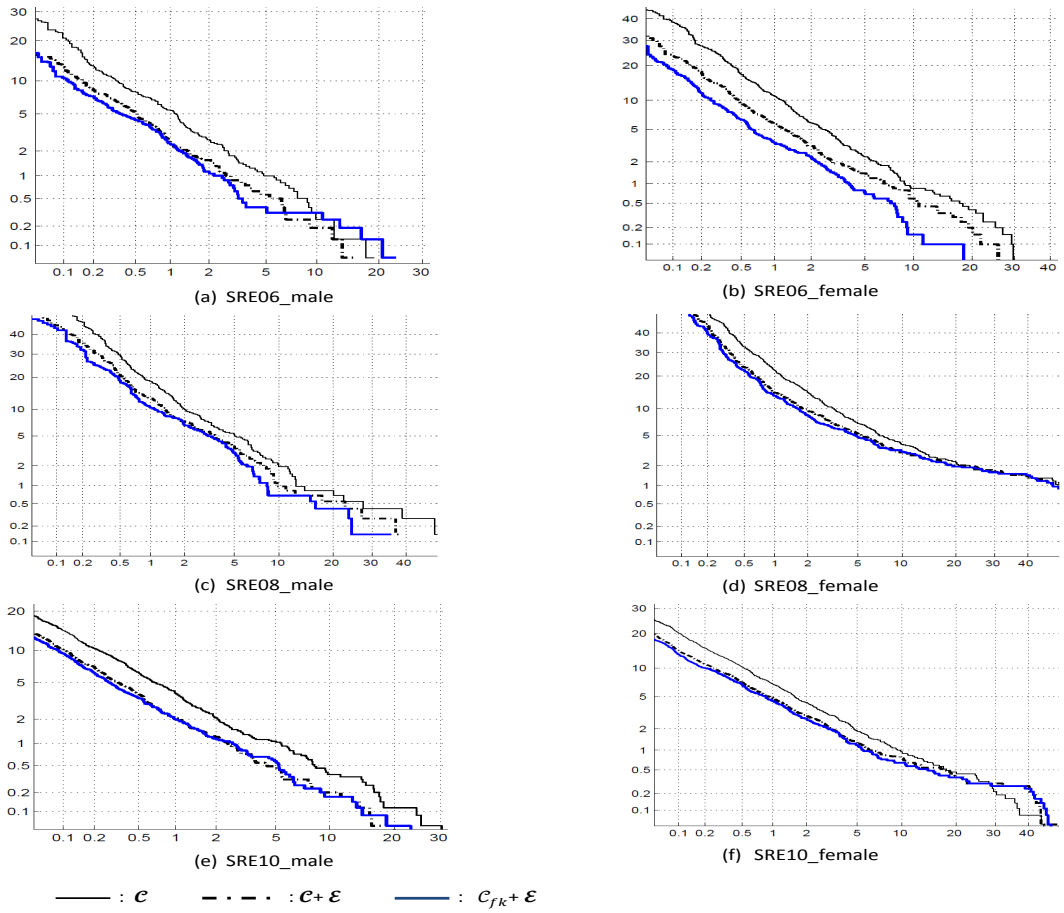

Figure 6: DET curves comparison of PLDA models trained by using different amount of data. The results are given for male and female trials of SRE06, SRE08 and SRE10. The x-axis shows *False Alarm Probability (in %)* and the y-axis shows *Miss Probability (in %)*.

Table 14: Results of SRE12(c2) using $P_{Known} = 1$ and $P_{Known} = 0$. For male, $k = 37$, and for female, $k = 25$.

Male model	$P_{Known} = 1$		$P_{Known} = 0$	
	C^{act}	C^{min}	C^{act}	C^{min}
C	0.340	**0.246**	0.341	0.340
$C_{a\text{-}k}$	0.339	0.257	**0.327**	**0.305**
$C_{a\text{-}fk}$	**0.336**	0.268	0.363	0.342
$\mathcal{R}$	0.348	**0.246**	**0.330**	0.337
$\mathcal{R}_{a\text{-}k}$	0.340	0.254	0.347	0.326
$\mathcal{R}_{a\text{-}fk}$	**0.334**	0.257	0.338	**0.314**
Female model	$P_{Known} = 1$		$P_{Known} = 0$	
	C^{act}	C^{min}	C^{act}	C^{min}
C	0.414	**0.239**	0.433	0.339
$C_{a\text{-}k}$	0.395	**0.239**	**0.380**	0.341
$C_{a\text{-}fk}$	**0.406**	0.240	0.414	**0.326**
$\mathcal{R}$	0.446	0.235	0.493	0.322
$\mathcal{R}_{a\text{-}k}$	**0.405**	0.235	**0.402**	**0.317**
$\mathcal{R}_{a\text{-}fk}$	0.424	**0.230**	0.452	0.322

fk-NN discarded only a few speakers. For female, k-NN reduced more data than fk-NN in most cases.

Table 15: Data reduction rate (%) by k-NN and fk-NN. In k-NN, k was optimized using SRE06. For male, $k = 37$, and for female, $k = 25$. M: Male model, F: Female model, m: reduction rate (%) of i-vectors and n: reduction rate (%) of speakers. For SRE12, the results refer to a-fk-NN and a-k-NN.

M	SRE06		SRE08		SRE10		SRE12	
	m	n	m	n	m	n	m	n
C_k	41.7	11.1	21.3	4.5	8.3	0.6	22.7	4.6
C_{fk}	55.2	19.5	28.6	7.1	25.8	4.2	31.8	7.7
$\mathcal{R}_k$	50.9	10.6	30.7	4.3	12.9	0.8	31.5	4.0
$\mathcal{R}_{fk}$	52.7	11.5	33.2	4.8	26.2	3.1	43.5	7.5
F	SRE06		SRE08		SRE10		SRE12	
	m	n	m	n	m	n	m	n
C_k	57.9	20.2	26.3	6.3	14.7	2.8	30.9	8.2
C_{fk}	57.9	20.2	23.0	5.8	15.7	3.0	11.5	1.9
$\mathcal{R}_k$	65.5	17.9	34.9	5.1	20.6	1.8	39.7	6.3
$\mathcal{R}_{fk}$	65.5	17.9	31.7	4.1	19.6	1.7	17.1	1.7

6. Conclusions

In this paper, we presented data selection methods for PLDA modelling, which is one of the state-of-the-art methods for i-vector scoring in text-independent speaker verification. Using k-NN we showed that we can choose a subset of the available

training data of the PLDA model, and improve the system performance for both male and female trials of SRE06, SRE08, SRE10, and SRE12. In order to avoid the difficulty of optimising k on a development set, we presented a robust way of selecting k, named fk-NN, which uses a *local distance-based outlier factor* (LDOF). This method discarded irrelevant or noisy training data of the PLDA model as much as the conventional k-NN without the need for tuning k. Using both k-NN and fk-NN, we achieved reduced EER and detection costs. We also proposed variations of these methods, including ik-NN which uses different k for different data points. We addressed issues such as the effect on unseen impostors, and the the robustness to noise.

Future directions are many. It would be interesting to see whether the performance of gender-dependent PLDA models can be improved by selecting data from the opposite gender. Our proposed data selection methods does not depend on any channel compensation techniques. Therefore, it would be a good idea to explore whether they can benefit from methods such as WCCN, NAP or LDA. We should also explore how much training data that is required for training an efficient PLDA model. Further developments of ik-NN, as well of schemes for adding discarded i-vectors from the selected speakers seem to be promising directions. Also, the current method uses the unique set of the selected i-vectors, and thus ignores the number of times the i-vectors have been selected. Taking this information into account could be interesting. We should also explore the effect of data selection by k-NN and fk-NN in GMM-supervector space. Its success may help us in reducing training time of the total variability matrix. However, in (Beyer et al., 1999), it has been argued that as the dimensionality increases, the distance to the nearest neighbour approaches the distance to the farthest neighbour. This holds true for a broad range of distributions and distance measures including cosine similarity measure (Radovanovic et al., 2010). Therefore, both k-NN and fk-NN using the cosine similarity or cosine distance may become ill-defined for high dimensional supervectors. Therefore, we need to explore other distance metrics.

References

Beyer, K. S., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When Is "Nearest Neighbor" Meaningful? In C. Beeri, & P. Buneman (Eds.), *ICDT* (pp. 217–235). Springer volume 1540 of *Lecture Notes in Computer Science*.

Biswas, S., Rohdin, J., & Shinoda, K. (2014). i-Vector Selection for Effective PLDA Modeling in Speaker Recognition. In *Odyssey* (pp. 100–105).

Brümmer, N. (2012). SRE'12 - BOSARIS Toolkit. In *https://sites.google.com/site/bosaristoolkit/sre12*.

Brümmer, N., & Villiers, E. d. (2010). The speaker partitioning problem. In *Odyssey* (pp. 194–201).

Campbell, W. M., Sturim, D. E., Reynolds, D. A., & Solomonoff, A. (2006). SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *ICASSP (1)* (pp. 97–100).

Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., & Dumouchel, P. (2009). Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *INTERSPEECH* (pp. 1559–1562).

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech & Language Processing*, *19*, 788–798.

Garcia-Romero, D., & Espy-Wilson, C. Y. (2011). Analysis of i-vector Length Normalization in Speaker Recognition Systems. In *INTERSPEECH* (pp. 249–252).

Graff, D., Canavan, A., & Zipperlen, G. (1998). Switchboard-2 Phase I. In *Linguistic Data Consortium, Philadelphia*.

Hasan, T., & Hansen, J. H. L. (2011). A Study on Universal Background Model Training in Speaker Verification. *IEEE Transactions on Audio, Speech & Language Processing*, *19*, 1890–1899.

Hasan, T., Lei, Y., Chandrasekaran, A., & Hansen, J. H. L. (2010). A novel feature sub-sampling method for efficient universal background model training in speaker verification. In *ICASSP* (pp. 4494–4497).

Hatch, A. O., Kajarekar, S. S., & Stolcke, A. (2006). Within-class covariance normalization for SVM-based speaker recognition. In *INTERSPEECH*.

Hermansky, H. (1990). Perceptual Linear Predictive (PLP) Analysis of Speech. *J. Acoust. Soc. Am.*, *57*, 1738–1752.

Huang, C.-L., & Ma, B. (2011). Maximum Entropy Based Data Selection for Speaker Recognition. In *INTERSPEECH* (pp. 2713–2716).

Ioffe, S. (2006). Probabilistic Linear Discriminant Analysis. In *ECCV (4)* (pp. 531–542).

Kanagasundaram, A., Vogt, R. J., Dean, D. B., & Sridharan, S. (2012). PLDA based speaker recognition on short utterances. In *Odyssey* (pp. 28–33).

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

Kenny, P. (2010). Bayesian speaker verification with heavy-tailed priors. In *Odyssey*.

Kenny, P., Boulianne, G., & Dumouchel, P. (2005). Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, *13*, 345–354.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *IEEE Transactions on Audio, Speech & Language Processing*, *15*, 1435–1447.

Mak, M. W., & Yu, H. B. (2010). Robust voice activity detection for interview speech in NIST speaker recognition evaluation. In *APSIPA ASC*.

McLaren, M., Baker, B., Vogt, R., & Sridharan, S. (2010). Data-driven background dataset selection for SVM-based speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, *18*, 1496–1506.

McLaren, M., Vogt, R., & Baker, S., B.and Sridharan (2009). Data-driven impostor selection for T-norm score normalisation and the background dataset in SVM-based speaker verification. *Advances in Biometrics*, *5558*, 474–483.

NIST (2006). The NIST Year 2006 Speaker Recognition Evaluation Plan. In *http://www.itl.nist.gov/iad/mig/tests/spk/2006/index.html*.

NIST (2008). The NIST Year 2008 Speaker Recognition Evaluation Plan. In *http://www.itl.nist.gov/iad/mig/tests/spk/2008/index.html*.

NIST (2010). The NIST Year 2010 Speaker Recognition Evaluation Plan. In *http://www.itl.nist.gov/iad/mig/tests/spk/2010/index.html*.

Pelecanos, J., & Sridharan, S. (2001). Feature Warping for Robust Speaker Verification. In *Odyssey* (pp. 213–218).

Prince, S. J. D., & Elder, J. H. (2007). Probabilistic Linear Discriminant Analysis for Inferences About Identity. *IEEE International Conference on Computer Vision*, (pp. 1–8).

Radovanovic, M., Nanopoulos, A., & Ivanovic, M. (2010). On the Existence of Obstinate Results in Vector Space Models. In *SIGIR* (pp. 186–193).

Senoussaoui, M., Kenny, P., Brümmer, N., de Villiers, E., & Dumouchel, P. (2011). Mixture of PLDA Models in I-Vector Space for Gender-Independent Speaker Recognition. In *INTERSPEECH* (pp. 25–28).

Sturim, D. E., & Reynolds, D. A. (2005). Speaker Adaptive Cohort Selection For Tnorm In Text-Independent Speaker Verification. In *ICASSP (1)* (pp. 741–744).

Suh, J.-W., Lei, Y., Kim, W., & Hansen, J. H. L. (2011). Effective background data selection in SVM speaker recognition for unseen test environment: More is not always better. In *ICASSP* (pp. 5304–5307).

Zhang, K., Hutter, M., & Jin, H. (2009). A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data. In *PAKDD* (pp. 813–822).