Int J Speech Technol (2014) 17:211–221
DOI 10.1007/s10772-013-9219-z

Effective background data selection for SVM-based speaker recognition with unseen test environments: more is not always better

John H.L. Hansen · Jun-Won Suh · Pongtep Angkititrakul · Yun Lei

Received: 16 August 2013 / Accepted: 21 November 2013 / Published online: 10 January 2014

Abstract This study focuses on formulating a procedure to select effective negative examples for the development of improved Support Vector Machine (SVM)-based speaker recognition. Selection of a background dataset, or a collection of negative examples, is the crucial step for building an effective decision surface between a target speaker and the non-target speakers. Previous studies heuristically fixed the number of negative examples used based on available development data for performance evaluation; nevertheless, in real applications this does not guarantee sustained performance for unseen data, as will be shown. In the proposed model selection framework, a novel ranking method is first exploited to rank order the negative examples for selecting a set of background datasets with various population sizes. Next, an error estimation and model-selection criterion are proposed and employed to select the most suitable target model among the model candidates. The experimental validation, conducted on the NIST SRE-2008 and SRE-2010 data, demonstrates that the proposed background data selection slightly but consistently outperforms the fixed-size background data selection, and achieves a relative improvement of +6 % over the non-selection background framework in terms of minDCF.

J.H.L. Hansen (✉) · J.-W. Suh · P. Angkititrakul · Y. Lei
Center for Robust Speech Systems (CRSS),
Erik Jonsson School of Engineering and Computer Science,
University of Texas at Dallas, Richardson, TX, USA
e-mail: John.Hansen@utdallas.edu

J.-W. Suh
e-mail: junwon.suh@utdallas.edu

P. Angkititrakul
e-mail: Pongtep.Angkititrakul@utdallas.edu

Y. Lei
e-mail: Yun.Lei@utdallas.edu

Keywords Speaker recognition · Support vector machine (SVM) · NIST SRE · Robustness in speaker ID

1 Introduction

Current state-of-the-art speaker recognition (Speaker Identification or SID) systems use either (i) i-vector system (Dehak et al. 2010), (ii) GSV-SVM: Gaussian supervector with a support vector machine backend (Campbell et al. 2006), or (iii) GMM-UBM: Gaussian mixture model with a universal background model (Reynolds et al. 2000). Many approaches also fuse these solutions in combination. So while i-vector approaches are currently very popular, studies have shown that fusing diverse systems can provide clear and consistent SID improvement. As such, one core state-of-the-art solution employs a Support Vector Machine (SVM) using a high- and fixed-dimensional supervector, which is obtained by concatenating all of the mean vectors of a Gaussian Mixture Model (GMM), as input feature (Dehak et al. 2010). In particular, only the adapted mean vectors of the universal background model (UBM) are exploited, while the covariances and mixing weights are shared among all speakers. Nevertheless, for a given target speaker, the supervectors estimated from different training utterances are subject to inter-session variability especially when these training samples come from different channels. As a result, they may not be properly scored against the trained speaker model. The factor analysis technique, particularly joint factor analysis (JFA), has been proposed to compensate for the variability of a Gaussian supervector as a linear combination of speaker and variability (i.e., channel traits/properties), where the SVM uses both knowledge sources as the SVM input features (Dehak et al. 2010; Kenny et al. 2007). Recently, such JFA compensation and Gaussian supervector

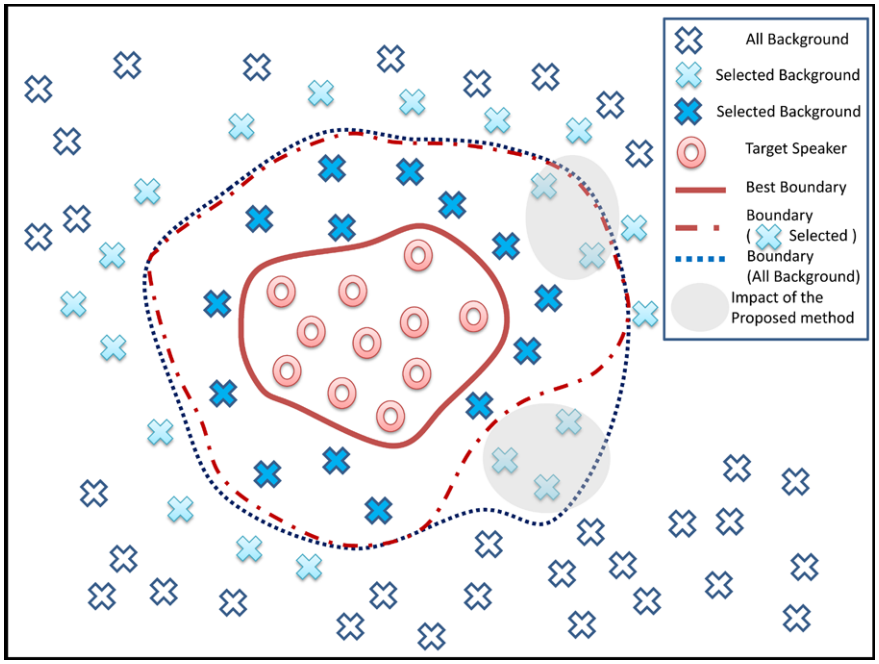

Fig. 1 A hypothetical illustration of background data selection for effective SVM boundary

SVM approaches have dominated the speaker recognition evaluation in NIST-SRE (Evaluation NSR 2008, 2010).

In general, for each target speaker, the SVM builds a hyperplane as the decision boundary between positive examples belonging to the target speaker and negative examples drawn from the impostor or background speakers (Joachims 1999). Here, the collection of negative examples is called the background dataset. The limited amount of positive examples for each target speaker requires effective selection to exploit the negatives examples, since in general there is an unlimited amount of the negative examples available. However, depending on microphones, handsets, communication channels, and other factors, not all impostor-speaker data is equally useful in building the best SVM. Hence, it is important to develop a framework that is capable of identifying those impostor speakers that are the closer cohorts to the target speaker, since employing too many speakers that are acoustically far from the target speaker does not contribute as much to an effective SVM hyperplane decision surface. Previous studies (McLaren et al. 2009, 2010) have shown that a flexible-size background dataset can improve performance of the SVM-based speaker recognition system, while classification actually degrades when using all available negative examples.

One of the drawbacks of employing an SVM background model is that it requires a specific number of negative examples to construct an effective hyperplane decision surface. Selecting an appropriate number of speakers in the background dataset is important since using too many negative examples introduces problems where the SVM will have a too narrow hyperplane, thus increasing the miss-rate for the overall system performance. On the contrary, selecting too small a number of negative examples makes the hyperplane excessively wide which would result in an increase in the false-alarm rate. Therefore, selecting the right amount of negative examples to build a proper hyperplane is important for training an effective speaker model using an SVM. Figure 1 hypothetically illustrates the idea of background data selection for a more effective (e.g., tighter) hyperplane decision surface.

Training an effective target speaker model from available background dataset requires the following three steps: First, it is necessary to rank order the negative examples in order to select the most suitable background dataset. Second, the target speaker model is trained using the positive examples against the negative examples belonging to the selected background dataset. It is possible that the target speaker models will have various sizes (i.e., number of SVM's support vectors) depending on the selected size of the background dataset. Third, an optimal model is selected among the potential target speaker models based on the estimated performance errors using a small amount of development data. In this study, we will focus on formulating effective and systematic ways to approach the first and third steps: ranking the background speakers and model selection.

Effective and efficient data ranking methods have been drawing broad attention in the research community recently. The training process for a ranking function plays a crucial role in extracting various types of information such as audio-based spoken dialogs, text documents, images, and video streams (Hansen et al. 2005; Joachims 2002; Cao et al. 2007; Lee et al. 2004). Various learning functions have also been proposed for information retrieval based on kernel classifiers (e.g., SVM) such as Pairwise (Joachims 2002) and List-

wise (Cao et al. 2007) methods. Each method uses a distinct learning function to rank order the objects in the set; the Pairwise approach considers object pairs as instances in learning, while the Listwise approach uses lists of objects as instances in its learning. In this current study, both Pairwise and Listwise approaches are introduced to rank order negative examples for training speaker SVMs. A discussion and comparison of both methods will be presented in the following sections.

After obtaining a collection of target model candidates using the selected background datasets, different error estimation and model-selection criteria are employed to select the most suitable model as a target model. Vapnik and Chapelle (2000) derived the expectation of the error boundary for the hyperplane using an SVM (Vapnik and Chapelle 2000; Duan et al. 2003), and these error estimation methods were used to tune the kernel parameters (Duan et al. 2003; Tuda et al. 2001; Chapelle et al. 2002). In this study, we study these error estimation methods and then propose a new error measurement to estimate errors of each trained speaker model, namely modified validation method, by splitting the development data into two distinct subsets and then estimating errors using the left-out subset from training. We will discuss and compare our proposed method with two previously proposed methods for error estimation, support vector count and radius margin, in the following sections.

For a new or unseen-data evaluation of speaker recognition systems, researchers generally fix the overall system parameters which include: (i) the entry number of the background dataset, (ii) the feature dimension, and (iii) the number of Gaussian mixtures (i.e., by using the available pre-evaluation data such as NIST SRE 2008 (Evaluation NSR 2008) for a new evaluation of NIST SRE 2010 (Evaluation NSR 2010)). It is clearly a challenging problem to find an optimal number of negative background entries of the dataset when the final speaker recognition system is deployed in a range of unknown conditions. Therefore, here we propose a system that employs an error estimation scheme to select the background evaluation dataset without tuning to a specific entry number for the dataset.

This paper is organized as follows. Section 2 describes the problem background and our motivation for selection of the background dataset. The methods of ranking background speakers are discussed in Sect. 3. Section 4 describes error measurement and model selection. The system description and specific parameter setting are included in Sect. 5. An extensive performance assessment along with results are presented in Sect. 6. Finally, conclusions are presented in Sect. 7.

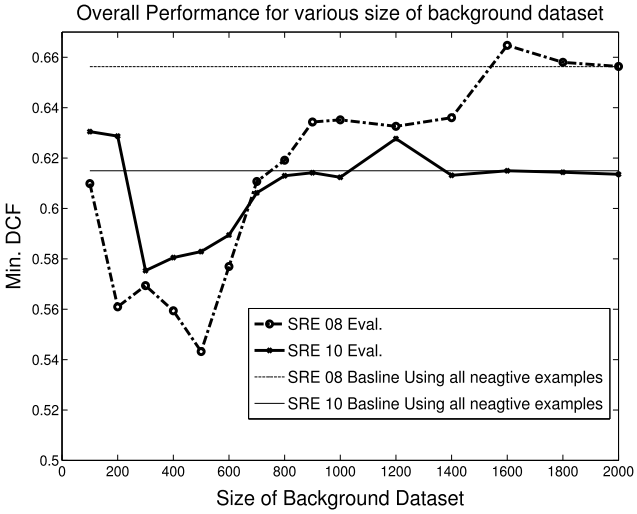

Fig. 2 MinDCFs of NIST-SRE 2008 and 2010 evaluation sets as the selected size of background dataset varies from 100 to 2000

2 Selection of background data

The idea of background dataset refinement for an SVM was previously introduced by McLaren et al. (2009, 2010). In their approach, the selection of the SVM background dataset was performed by ranking the candidate examples using a suitability metric called the Support Vector Frequency, which is the total number of instances a particular vector is selected as a support vector when training a set of development target models. Subsequently, a collection of datasets, each containing different sizes of examples with the highest support vector frequency, is used to determine the best number of examples. However, a fixed number of negative examples does not always guarantee a consistent best performance with a new evaluation data. To illustrate this point, we evaluated a speaker recognition system using the NIST SRE-08 and SRE-10 data (i.e., male 5 min), where the background data was drawn from the NIST SRE-04 and SRE-05. More details regarding this speaker recognition system will be further discussed in the following sections. Figure 2 shows the results of this evaluation in terms of minDCF (i.e., minimum Detection Cost Function (Evaluation NSR 2008, 2010) for both evaluation sets when the size of the background dataset varies from 100 to 2000, as well as using all available background data. It is obvious that exploiting all available background data does not guarantee the best performance. Furthermore, we can observe that the background dataset with 500 negative examples gives the best performance when using the SRE-08 evaluation set. However, this same size of negative examples does not give the best performance when testing with the SRE-10 evaluation set, which can be achieved when only 300 negative samples are selected instead. As the size of negative samples increases, the performance of both evaluation sets tends to converge to

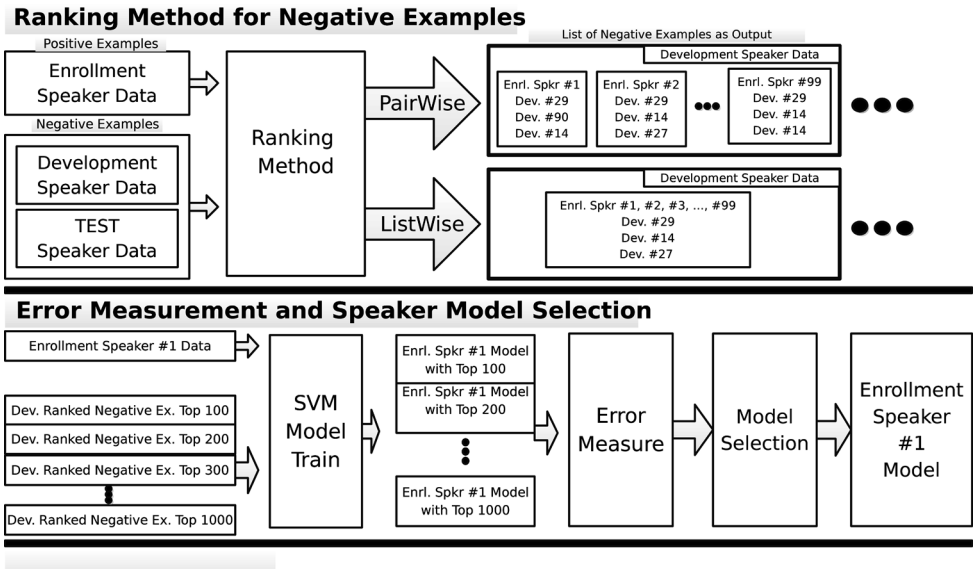

Fig. 3 Overall block diagram of the proposed system

the baseline performance that uses all available background datasets. These variation of performances across different datasets suggests that an SVM speaker recognition system needs an improved method for finding the optimal size and content of the effective background datasets.

The following sections introduce the proposed background dataset selection method. The main objective of our method is to provide a unified, data-driven approach to automatically select the optimal size of the background dataset for different evaluation sets without pre-assigning a specified size for background data. To accomplish this, a ranking method is introduced for finding the most informative negative examples. Next, an error measurement and model selection procedure for the background dataset are employed to select the best set. These steps represent the proposed SVM speaker selection process, and are summarized in the flow diagram in Fig. 3. In the following sections, we will describe these two steps in details.

3 Ranking the background speakers

The procedure for effective background dataset selection starts by identifying the best negative examples from a virtually unlimited amount of data. A structured selection method can be applied to the selection of a set of negative examples. In this study, we considered two ranking methods: Pairwise and Listwise, where the learning methods are based on object pairs and lists of objects, respectively.

3.1 Pairwise approach

The pairwise method is similar to a general binary classification process between two categories, where a ranking function is formulated to minimize the empirical risk function (Joachims 2002) of a binary ordering relation, not a class label. The Pairwise method focuses on the relevance of each object (i.e., target speaker) to build the best overall SVM speaker model that rejects the non-target speakers. The ranked results are sorted in a descending order, which are arranged from the most relevant to the least relevant target objects. Here, we utilized the SVMLight toolkit which has a ranking function that scores the relevance of the negative examples compared to each target object (Joachims 1999). The pairwise approach can be applied straightforwardly and the object pairs can be easily obtained in most scenarios. However, the pairwise approach is formulated to minimize the classification errors of object pairs, rather than minimizing the ranking errors of target speakers.

3.2 Listwise approach

The Listwise loss function incorporates the scores or frequency of occurrence information into the results of the Pairwise method, and the ranking function is trained so as to maximize the Listwise loss function (Cao et al. 2007). In this study, the Listwise loss function uses the frequency of occurrence from the most relevant to least relevant negative examples that have appeared at least twice across all the target speakers, and the results of this function are used for the background dataset.

Figure 4 pictorially illustrates the concept of both ranking methods, where the target speakers are represented by triangles (A, B, and C), and the small numeric circles ($1, 2, 3, \ldots$) represent the background speakers. Subsequently, d_{A-1} is the distance between target speaker A and background speaker 1, and the d_{ABC-1} represent the Listwise distance between all target speakers and background

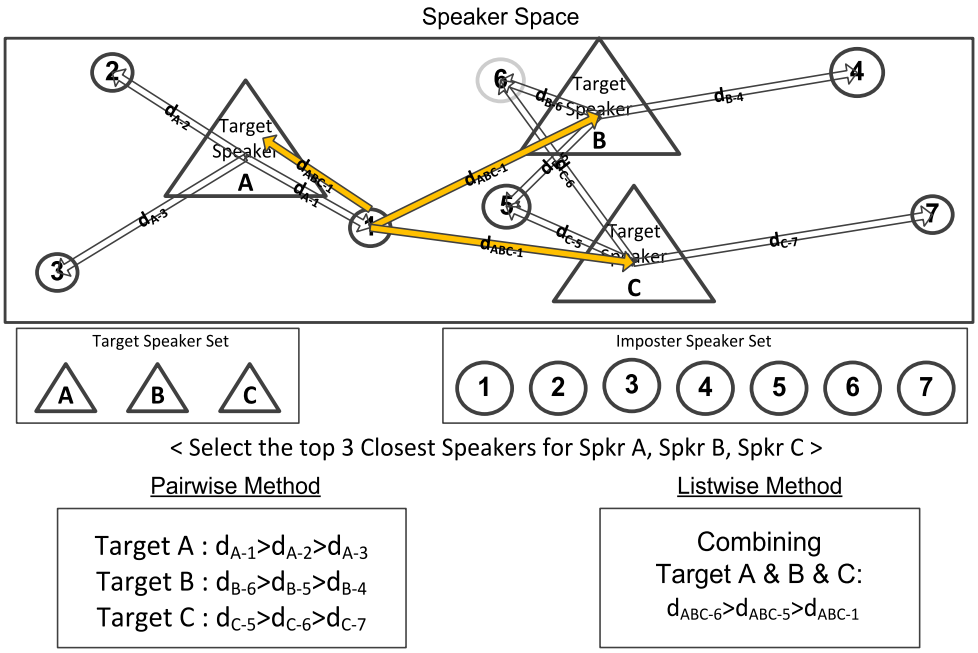

Fig. 4 Pictorial illustration of the Pairwise and Listwise ranking methods

speaker 1. Similar annotation is applied to other distance pairs between target and background speakers. As a result, employing the Pairwise approach to the target speaker *A* will rank the top-3 background speakers as $1 \rightarrow 2 \rightarrow 3$. For the Listwise approach, the highest ranked background speaker considering all target speakers is the background speaker 6, since it ranks in the top-2 for the target speakers *B* and *C* with the total distance lower than the background speaker 5, which also ranks in the top-2 for both target speakers *B* and *C*.

3.3 Comparison of ranking approaches

The advantage of employing a ranking method for background data selection is shown in Fig. 5, where we compare two methodical selection approaches, *Pairwise Method* and *Listwise Method*. For completeness, we also compare their results with randomly selected background speakers of the same population size (e.g., Random). The experiments were conducted using data from the SRE-08 evaluation set. Both ranking methods show a better system performance than that of the Random Selection when less than 1200 negative examples were utilized, and the Listwise Method showed superior performance over the Pairwise Method as the background dataset is reduced. Further analysis reveals that the Pairwise Method ranking system does not construct the speaker model effectively, since the most relevant background speakers for each target speaker contribute only to building the hyperplane. That is, a hyperplane built by the Pairwise Method is so narrow that it will not only reject all impostor speakers, but also reject many of the correct test examples. The key finding here is that any further amount

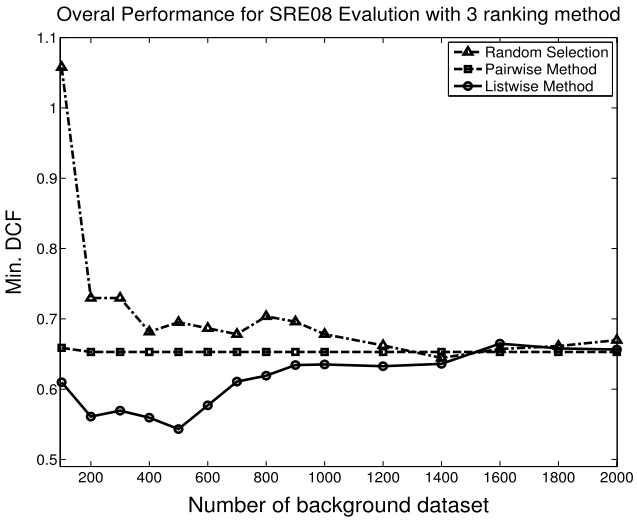

Fig. 5 Comparison of Pairwise and Listwise ranking methods for background data selection, as well as random selection, using NIST-SRE 08 evaluation set

of background speakers do not provide improved discriminative power for a target speaker in an SVM framework. The Pairwise Method, therefore, does not provide as useful information for constructing the effective SVM hyperplane when the negative examples exceed 200. Alternatively, the Listwise Method provides more consistent performance, since the various background speakers that are close to all the target speakers would provide more effective discriminating information in building an effective decision hyperplane.

4 Error measurement and model selection

For an SVM-based speaker model, a hyperplane is trained as a decision boundary between a target speaker and a pool of ranked background speakers. In order to assess the effect of the size of the selected background population, a collection of candidate sizes of ranked background speakers is used to construct different SVMs. In this study, we varied the size of the background dataset, p, with an increment of 100, (i.e., $p \in \{100, 200, 300, \ldots, 900\}$).

For each background dataset p, the trained model consists of the support vectors, $\mathbf{x}_i$, a norm vector, $\mathbf{w}_p$, and a bias, b, that satisfies the following inequality:

$$y_i(\mathbf{x}_i \cdot \mathbf{w}_p + b) - 1 \geq 0 \quad \forall\, i \in y_i = \pm 1, \tag{1}$$

where y_i is either 1 or -1 indicating the class labels (target speaker versus background speakers), $i = 1, \ldots, K$, and K is a total number of training entries. The norm vector, $\mathbf{w}_p$, is formulated to maximize the space between both the positive and negative classes (Burges 1998), as

$$\mathbf{w}_p = \sum_i \alpha_i y_i \mathbf{x}_i. \tag{2}$$

The output function for the kth example of the dataset p is therefore expressed as,

$$O_{p,k} = \mathbf{w}_p \cdot \mathbf{x}_k - b. \tag{3}$$

Having briefly presented the basic SVM framework, we now discuss the error measurement for assessing speaker models.

4.1 Error measurement

Here, the error is assessed by using three alternate schemes. The first two methods are derived from a mathematical analysis by Vapnik for obtaining an upper error bound using a Leave-One-Out (LOO) estimation (Vapnik and Chapelle 2000; Chapelle et al. 2002),[1] and the last method is our proposed method, termed the "modified validation method."

4.1.1 Support vector (SV) count method

This method directly exploits *support vectors* which is a subset of both positive and negative examples that are used to define the separating hyperplane in the SVM training stage. These support vectors, therefore, are the most informative examples that are employed in the evaluation stage (Chapelle et al. 2002). The upper error bound for the LOO is given as:

$$Err_{SV} \leq \frac{N_{SV}}{l}, \tag{4}$$

where N_{SV} is the total number of support vectors and l is the total number for the background dataset.

4.1.2 Radius-margin method

This method estimates the upper bound on the amount of error for the LOO scheme (Vapnik and Chapelle 2000) as,

$$Err_{RM} \leq \frac{1}{l}\frac{R^2}{\gamma^2},$$

where the margin γ and the radius R are defined by

$$R = \min_{\mathbf{a},\mathbf{x}_i} \|\mathbf{x}_i + \mathbf{a}\|$$

$$\gamma = \min_{\mathbf{x}_i, y_i} \frac{y_i(\mathbf{x}_i \cdot \mathbf{w}_p + b)}{\|\mathbf{w}_p\|},$$

such that R is the radius of the smallest sphere that contains all $\mathbf{w}_p$ vectors. In this study, the denominator of both error estimation methods is removed, since the increment of 100 negative examples makes comparisons between the models intractable.

4.1.3 Modified validation method (MVM)

This method is used to estimate the errors by splitting the development data into two distinct subsets (Duan et al. 2003), and then estimating errors using the left-out subset from training as a test set. In particular, the proposed method uses the top 900 ranked negative examples as a training subset and the remaining as the test subset. Note that the test subset is less relevant to all the target speakers than the training subset since all the negative examples are ranked in order by the employed ranking method. The modified validation error measurement counts the errors of the target speaker model using the same threshold and the same test subset data. The error measurement of the target speaker of population size n with the background dataset of size p is then defined as:

$$Err_{n,p} = \frac{1}{m} card\{j : O_{p,j} - \theta < 0\}, \tag{5}$$

where m is the total number of test subset examples, and $card\{\}$ is the cardinality of the set. The test subset example, j, is evaluated using Eq. (3). When the output of example j is less than a pre-defined threshold, the number of examples is counted as error. The size of the background dataset is varied from 200 to 900. The decision threshold θ is obtained by using the first $p = 100$ background dataset, since

[1] The leave-one-out (LOO) procedure consists of excluding one example from the training data, constructing the decision rule from the remaining training data, and then testing on the removed example. The process can be applied to every example of training data.

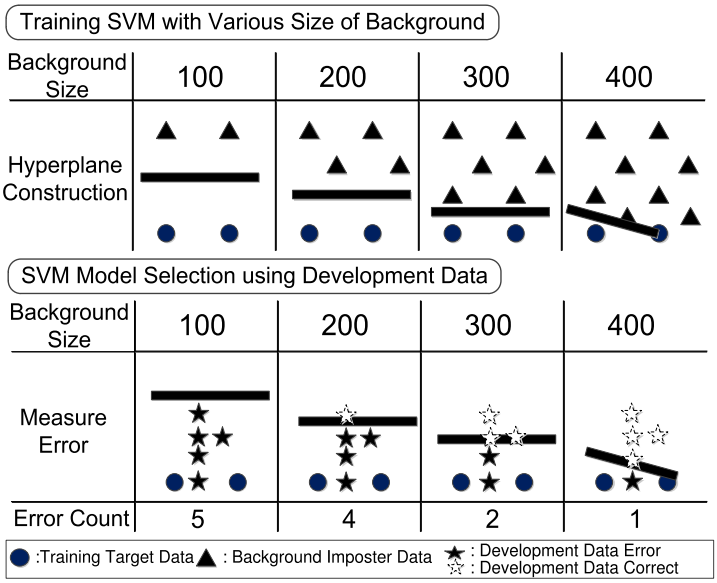

Fig. 6 Idea of selection of target-speaker model using the proposed modified validation method. The impostor development data represented as *stars* are moved from solid black entries, incorrectly labelled as target speakers, to above the decision plane, resulting in correctly rejected impostors as *white stars*

it contains the most relevant negative examples compared with the positive examples, as

$$\theta = \frac{1}{100}\sum_{k=1}^{100} O_{p,k}. \tag{6}$$

The overall idea of the modified validation method is pictorially illustrated in Fig. 6. The upper row of plots show the boundary hyperplanes obtained by training different sizes of background dataset (i.e., $p = 100, 200, 300, 400$). As the background population increases, the decision space margin between the positive (shown as circles) and the negative examples (shown as triangles) is expected to decrease. In the corresponding lower row of plots, the first 100 negative examples (i.e., development data) are used to set a threshold θ such that these 100 negative examples are initially categorized as errors. However, as the decision hyperplanes becomes tighter with increasing p count, the more data (e.g., negative data as stars) are correctly classified.

For completeness, Fig. 7 shows plots of the average error measure for speaker recognition evaluation employing the three error measurement strategies versus an increasing background dataset size from 200 to 900. The results are shown for both SRE-08 and SRE-10 datasets, with background datasets drawn from the SRE-04 and SRE-05 datasets. As can be seen, the error profiles all decrease with an increase in background population size. In the next section, a procedure is developed to select the best model for each set of target speakers.

4.2 Model selection using error difference

The difference between two errors (defined in Eq. (5)) generated by two speaker models trained from different background datasets is called the Error Difference (*ErrDiff*), and is defined as follows,

$$ErrDiff_{n,p} = |Err_{n,p} - Err_{n,p+\Delta}|, \tag{7}$$

where p represents the size of the background dataset, and the increment Δ is set to 100 in this phase of the study. This difference of estimated errors reflects the intrinsic migration of the SVM decision hyperplane as the size of negative dataset is increased. The averages of *ErrDiff* employing three different error measurements are shown in Fig. 8. The optimal number of background speakers resulting in the best performance previously seen in Fig. 2 is correlated with the slope of the *ErrDiff* plot in Fig. 8(C). Here, the best performance for SRE-08 can be achieved using a set of 500 background speakers, where the steepest slope also occurs at the dataset of size 500. A similar trend is also observed for the SRE-10 evaluation. The background dataset selection is performed based on the steepest slope of the *ErrDiff* for each of the target speakers, as

$$\begin{aligned} p^* &= \arg\max_p \{ErrDiff_{n,p} - ErrDiff_{n,p+\Delta}\}, \\ p &= 200, \ldots, 800, \end{aligned} \tag{8}$$

where $ErrDiff_{n,p}$ is the model to predict the least *ErrDiff* between the two models; therefore the $p + \Delta$ background dataset is selected to train the SVM for the target speakers. Again, the error is decreasing with an increment of dataset size in a similar manner to that shown in Fig. 7.

5 System description

In this section, the system description for experimental evaluations using the SVM data selection solutions is presented.

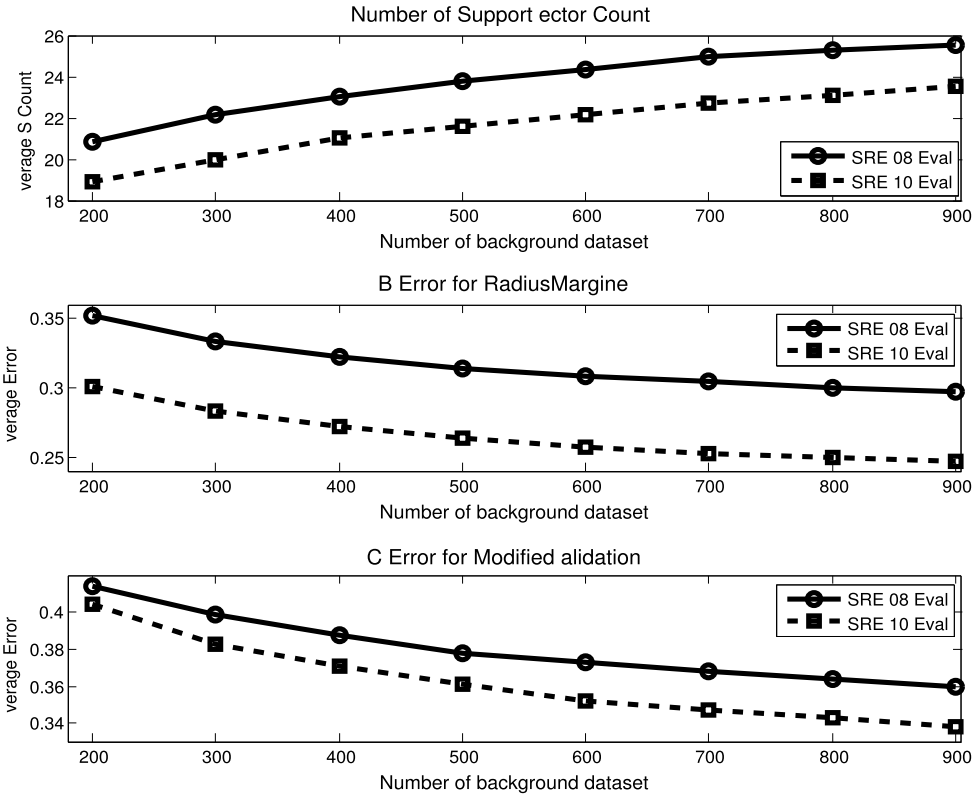

Fig. 7 Average error measure obtained from three different error measurements ((**A**) Support Vector Count, (**B**) Radius Margin, (**C**) Modified Validation Method (MVM)) with increasing number of background dataset size from 200 to 900 speakers

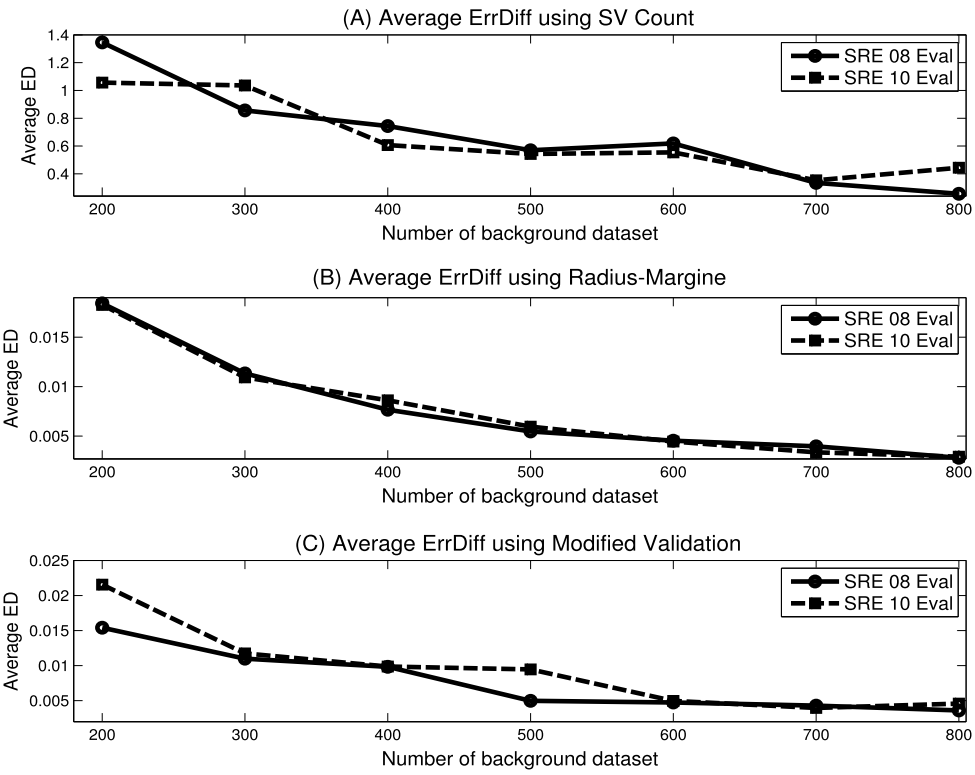

Fig. 8 Average Error Difference (*ErrDiff*) of three different error measurements as the size of background dataset increases

The evaluation results will be discussed in the following section.

5.1 Baseline SVM system

For parameterization, a 60-dimensional feature set (19 MFCC with log energy $+\Delta + \Delta\Delta$) was used, where features were extracted using a 25 ms analysis window with a 10 ms skip rate, filtered by feature warping with a 3-s sliding window. The system also employed Factor Analysis (Kenny et al. 2007), followed by Linear Discriminative Analysis (LDA) and Within Class Covariance Normalization (WCCN) (Dehak et al. 2010) for the SVM system. Similar SVM processing have also been employed in previous studies (Dehak et al. 2010) which represents our baseline here. Next, the NIST SRE-2004, 2005, 2006 enrollment data were used to train gender-dependent UBMs with 1024 mixtures. The total variability matrix was trained on the Switchboard II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST SRE-2004, SRE-2005, and SRE-2006 male enrollment data with 5 or more recording sessions per speaker. A total of 400 factors were used. The LDA matrix was trained on the same data as that we used for construction of the total variability matrix. In our experiments, the dimension of the LDA matrix was set to 140. Finally, the within class covariance matrix was trained using NIST SRE-2004, and SRE-2005 data, and a cosine kernel was used to build the SVM systems.

5.2 Evaluation dataset

The proposed algorithm was evaluated on the 5 min-5 min telephone-telephone condition of the NIST 2008 and 2010 speaker recognition evaluation (SRE) corpora (Evaluation NSR 2008, 2010), (i.e., SRE-08 and SRE-10). The evaluation dataset was limited to male speakers.

5.3 Background dataset

The background dataset consists of NIST SRE-04, and SRE-05 with a total of 2718 utterances. Each of the utterances is parameterized as previously discussed and then is exploited as a negative example. The Listwise method is employed to rank all 2718 negative examples, and the top 100 negative examples are used to set the decision rule. The target speaker models are built from 200 to 900 negative examples, and the remaining 1718 examples are used as development data for error estimation.

6 Experimental results

In current investigations, the speaker-recognition community has exploited the detection cost function (DCF) and Equal Error Rate (EER) as standard measurements to assess overall system performance. Therefore, our primary goal here is that the proposed background dataset selection method should achieve an equivalent or improved system performance in terms of both DCF and EER, compared with the previous studies (McLaren et al. 2009, 2010) without presetting the dataset size with a fixed number. Hence, the criterion of success should be no major change in performance between the expected level of performance and what is actually achieved with the unseen test data.

Table 1 The number of target speakers employing each background dataset

Background Dataset Size	300	400	500	600	700	800	900
SRE-08	175	144	129	69	57	45	29
SRE-10	347	265	182	187	91	69	62

Table 2 Results of SRE-08 and SRE-10 evaluation employing three different error measurements: SV Count, Radius-Margin, and Modified Validation, compared with the baseline system without background selection

	SRE-08		SRE-10	
	Min. DCF	EER	Min. DCF	EER
Baseline	0.656	6.01	0.614	6.14
SV Count	0.617	5.67	0.582	5.83
Radius-Margin	0.587	5.38	0.585	5.84
Modified Validation	0.542	4.98	0.547	5.50

6.1 Background dataset selection analysis

A diverse background dataset is selected for each target speaker using the proposed *ErrDiff* method. Table 1 summarizes the number of target speakers sharing the same size of selected background dataset for both NIST SRE-08 and SRE-10 evaluation datasets using the modified validation error method. As can be seen, most target speakers require a smaller sized background dataset with the majority around 300–400 negative examples. Flexible selection of the background dataset size helps the system focusing on the hyperplane near the target speaker and thereby improves the effectiveness of the background dataset.

6.2 Evaluation of different error estimation schemes

Table 2 summarized the results of SRE-08 and SRE-10 evaluations with the *ErrDiff* computed from three different error measurements discussed earlier: (i) SV count, (ii) Radius Margin, and (iii) Modified Validation. For comparison, the baseline represents a system which employs all

Table 3 The proposed method result comparing the best 500 entries of background dataset set for SRE-08

	Fixed Number Background Dataset			Proposed *ErrDiff* Selection Method	
	Fixed Number	Min. DCF	EER	Min. DCF	EER
SRE-08	500	0.543	4.98	0.542	4.97
SRE-10	500	0.583	5.83	0.547	5.49
SRE-10	300	0.575	5.79		

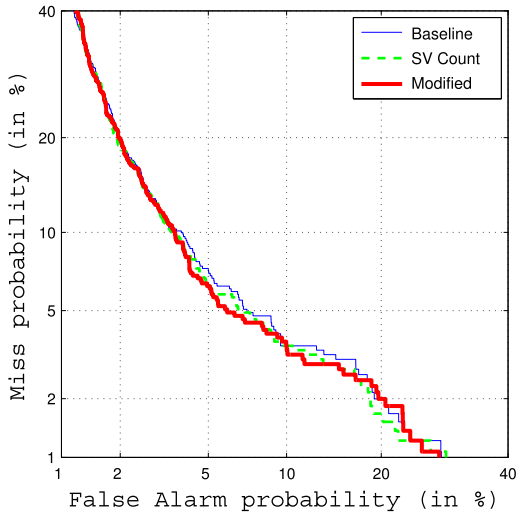

Fig. 9 DET curves of SRE-08 evaluation using the proposed Modified Validation, compared with the SV count and the baseline system

available background speakers. All background data selection approaches resulted in better performance than the baseline system in terms of both EER and minDCF. The Radius-Margin and Modified Validation methods provided a consistent performance gain for both SRE-08 and SRE-10 evaluations, but the SV count method did not perform well on the SRE-08 data. The proposed Modified Validation method consistently outperformed the other two error measurements. Therefore, this method is used for estimating errors of background data selection. For completeness, Fig. 9 compares the Detection Error Tradeoff (DET) curves of the SRE-08 results for the Modified Validation and SV count error measurements, compared with the baseline system. The SV count and Modified Validation schemes both out perform the baseline especially near the EER point.

6.3 Background dataset selection evaluation

In this section, we compare performance of the proposed *ErrDiff*, where the optimal size of the background dataset varies from speaker-to-speaker, with the fixed-size background selection. For the fixed-number background dataset, we used the best result of the background dataset selected by the ListWise ranking method (i.e., 500 negative examples for SRE-08, 300 negative examples for SRE-10). Table 3 shows the results of the best background dataset selected by the Listwise Ranking method and our proposed *ErrDiff* selection method. Again, 2718 utterances from the SRE-04 and SRE-05 are used as the background dataset, and subsequently 15 candidate background datasets are evaluated as potential background datasets for the baseline system. For each background dataset, the background size is incremented by 100 until it reaches 1000, and then by 200 until it reaches 2000 to represent the negative examples in the dataset, as previously shown in Fig. 2. The proposed *ErrDiff* selection method uses 8 distinct background datasets, incrementally by 100 from 200 up until 900, and separate examples are used for the error estimation calculation. For SRE-08, a fixed-size background dataset of 500 gave the best result of minDCF equal to 0.543, while the proposed *ErrDiff* selection method gave a slightly better level of performance with minDCF equals 0.542. Subsequently, the background dataset of size 500 is also applied to the SRE-10 evaluation, and the proposed method also shows an improve minDCF performance of 0.547, resulting in a +6 % relative improvement from the best result. The best performance for SRE-10 data with a fixed background number of 300 is a minDCF of 0.575, while the proposed method outperforms this with a minDCF of 0.547. This highlights the consistency of the proposed *ErrDiff* selection process in ensuring performance for unseen test data.

7 Discussion and conclusions

There is consensus in the speaker ID community that a fusion of sub-systems such as: (i) i-vector, (ii) GSV-SVM (Gaussian Supportvector with SVM backend), and (iii) GMM-UBM can provide complimentary strengths and improve overall system performance. Therefore, effective data selection for SVM speaker ID remains an important research challenge. In this study, a new method was proposed to find the best background dataset for SVM construction without fixing a number of negative examples for every speaker model. The use of a novel ranking method to rank the candidate negative examples, and the criterion of the most *ErrDiff* difference is used to select the most suitable background dataset for each target speaker. This background dataset is then used as the negative examples for training the target speaker model. In this manner, target speakers are trained with the most effective informative and flexible size of negative speaker examples.

Experimental validations with a pool of background speakers drawn from the NIST SRE-04 and SRE-05 datasets showed that the selection of the background dataset using

the *ErrDiff* method resulted in the best performance in terms of both minDCF and EER. The selection of the background dataset using *ErrDiff* is also more robust for new unseen data than selecting a prior fixed number dataset. The proposed method also enables the resulting system to reach the minDCF for a new background dataset or evaluation data.

For future work, a range of alternative kernels could be studied to project the support vectors into higher dimensions. Alternative ranking methods and accurate expectation of the error bound of the SVM can also be studied to select an even more effective background dataset, perhaps conditioned on fixed computing resources. An automatic measurement of the percentage used for the bottom test evaluation dataset could also help in further establishing a fully automatic system configuration. This work has therefore established a speaker model selection and score normalization process that provides both effective and consistent performance for speaker recognition.

Acknowledgements This project was funded by USAF under FA8750-12-1-0188, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

References

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

Campbell, W., Sturim, D., & Reynolds, D. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308–311.

Cao, Z., Qin, T., Liu, T., Tsai, M., & Li, H. (2007). Learning to rank: from pairwise approach to listwise approach (pp. 129–136). ACM.

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, *46*, 131–159.

Dehak, N., Kenny, P., Dehak, R., Ouellet, P., & Dumouchel, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Duan, K., Keerthi, S., & Poo, A. (2003). Evaluation of simple performance measures for tuning SVM hyperparameters. *Neurocomputing*, *51*, 41–59.

Evaluation NSR (2008). The NIST year 2008 speaker recognition evaluation plan.

Evaluation NSR (2010). The NIST year 2010 speaker recognition evaluation plan.

Hansen, J., Huang, R., Zhou, B., Seadle, M., Deller, J., Gurijala, A., Kurimo, M., & Angkititrakul, P. (2005). Speechfind: advances in spoken document retrieval for a national gallery of the spoken word. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 712–730.

Joachims, T. (1999). *SVMLight: support vector machine*. http://svmlight.joachims.org/, University of Dortmund.

Joachims, T. (2002). Optimizing search engines using clickthrough data (pp. 133–142). ACM.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Speaker and session variability in GMM-based speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(4), 1448–1460.

Lee, M., Keerthi, S., Ong, C., & DeCoste, D. (2004). An efficient method for computing leave-one-out error in support vector machines with Gaussian kernels. *IEEE Transactions on Neural Networks*, *15*(3), 750–757.

McLaren, M., Vogt, R., Baker, B., & Sridharan, S. (2009). Data-driven impostor selection for T-norm score normalisation and the background dataset in SVM-based speaker verification. *Advances in Biometrics*, *5558*(7), 474–483.

McLaren, M., Vogt, R., Baker, B., & Sridharan, S. (2010). Data-driven background dataset selection for SVM-based speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1496–1506.

Reynolds, D., Quatieri, T., & Dunn, R. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10*, 19–41.

Tuda, K., Rätsch, G., Mika, S., & Müller, K. (2001). Learning to predict the leave-one-out error of kernel based classifiers. *Artificial Neural Networks*, *2130*, 331–338.

Vapnik, V., & Chapelle, O. (2000). Bounds on error expectation for support vector machines. *Neural Computation*, *12*(9), 2013–2036.