# A FAST SPEAKER VERIFICATION WITH UNIVERSAL BACKGROUND SUPPORT DATA SELECTION

*Gang Liu, Jun-Won Suh, and John H.L. Hansen*[*]

CRSS: Center for Robust Speech Systems
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas 75083, USA
{gxl083000, jxs064200, John.Hansen}@utdallas.edu

## Abstract

In this study, a fast universal background support imposter data selection method is proposed, which is integrated within a support vector machine (SVM) based speaker verification system. Selection of an informative background dataset is crucial in constructing a discriminative decision super-plane between the enrollment and imposter speakers. Previous studies generally derive the optimal number of imposter examples from development data and apply to the evaluation data, which cannot guarantee consistent performance and often necessitate expensive searching. In the proposed method, the universal background dataset is derived so as to embed imposter knowledge in a more balanced way. Next, the derived dataset is taken as the imposter set in the SVM modeling process for each enrollment speaker. By using imposter adaptation, a more detailed subspace per target speaker can be constructed. Compared to the popular support-vector frequency based method, the proposed method can not only avoid parameter searching but offers a significant improvement and generalizes better on the unseen data.

**Index Terms**: speaker verification, universal background dataset selection, adaptation, SVM, UBS

## 1. Introduction

In recent years, state-of-the-art speaker verification systems usually employ a support vector machine (SVM) along with Joint Factor Analysis (JFA)-based features as input. Most studies dedicated to SVM-based speaker verification have focused on optimizing performance through novel kernel design and tuning of the associated parameters [1,3]. However, other factors, especially the selection of background imposter data, can also impact the SVM decision hyper-plane positioning and in turn the classification performance.

Several recent studies on imposter data selection have shown promising results but with some limitation [4,5]. Some perform background dataset selection based on knowledge of the broad characteristics expected in the evaluation imposter trials such as language, age, gender and the audio source. Data sources that can maximize the matched condition will be employed to compose various background dataset to characterize the expected imposter space in the evaluation phase. Good performance can be obtained with this heuristic background dataset selection, however, the emphasis on the general information of the *entire* set may overlook the information embedded in each *individual* background imposter. One solution is to zoom in on the individual imposter and find the subset of the entire background dataset according to some imposter candidate fitness measurement. This is called the data-driven method. With this method, the optimal parameter configuration derived from development data (such as NIST SRE2008, for short, NIST08) is then applied to the evaluation data (such as NIST10). These parameters do not guarantee sustained performance for new unseen evaluation data. Also, the potential benefit is acquired at the cost of expensive and slow parameter searching.

The proposed method first makes use of all available enrollment and imposter speakers' data to build a single SVM, from which a universal imposter background dataset is derived. This is a more balanced imposter selection method with respect to the entire enrollment speakers' space, which, theoretically, can avoid the over-fitting issue associated with other methods.

This paper is organized as follows. Sec. 2 describes the baseline and popular system. Sec. 3 describes the proposed method for effective dataset selection. The system description and specific parameter setup are detailed in Sec. 4. A comprehensive performance assessment and results are given in Sec. 5. Finally, research finding are summarized in Sec. 6.

## 2. Baseline and Support Frequency Method

The SVM is a discriminative classifier trained to separate classes. The low-dimensional input vectors are usually projected into high-dimensional space where a separating hyperplane is positioned to maximize the margin between the labeled data [6]. Once the SVM training is done, the classifying hyperplane structure of the high-dimensional space is captured with a small subset of both positive and negative samples from the training dataset, which are termed as *support vectors*. Samples that are selected as the support vectors shared a common property of being the most difficult to classify as they lie on or within the margin between classes. In contrast, those data samples that are not selected as support vectors provide no clue in hyperplane construction. By the terminology of SVM, the normal of the classification hyperplane is given by $\omega = \sum_j \alpha_j y_j x_j$ where $x_j$ is the $j^{th}$ sample with class label $y_j \in \{-1, +1\}$ and $\alpha_j$ is the coefficient assigned to the $j^{th}$ example. Those examples with a positive coefficient are defined as support vectors/examples, while vectors/examples with zero coefficients have no impact on the final positioning of the hyperplane. The SVM model training is nothing but the process of selection of a subset of support vectors from the samples and weighting properly according to individual impact.

[*]This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

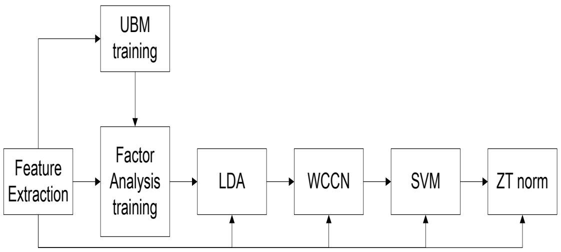


Figure 1. *Baseline speaker verification system block diagram*

Based on this observation, the support vector frequency of an example provides a measure of its relative importance in the background dataset. The support vector frequency ($SVFreq_j$) of the $j^{th}$ imposter example is defined as Eq. 1[4]:

$$SVFreq_j = \sum_{k=1}^{K}\delta_j^k \qquad (1)$$

where $\delta_j^k$ is 1 only if the $j^{th}$ imposter is a support vector in the $k^{th}$ model, and 0 otherwise. $K$ is the number of enrollment speakers. That is, the support vector frequency of an example is defined as the number of times that vector is selected as a support vector while training a set of SVMs on a development dataset. The resolution of the support vector frequency metric is dependent on the size of development dataset.

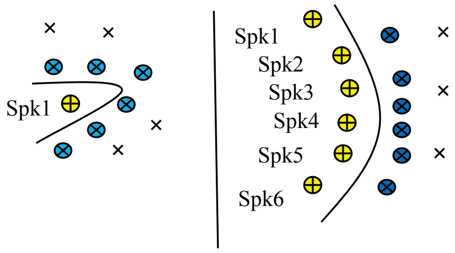
Following Eq.1, we can select the top $N$ most frequent support imposter vectors (or negative support vector) and this method is employed by many NIST SRE10 sites [7] and is called the Support Vector Frequency (*SVFreq*) method, where $N$ is a parameter need to be searched. Since we have reservations concerning the efficiency of this method, we will still use the traditional system without any imposter data selection as our baseline.

## 3. Proposed Method

We note that during the SVM modeling process, only the support vector has impact on the final classification performance, and we also note the fact of building an SVM with only one positive example (enrollment speaker) and thousands of negative example (imposter speaker) need to carefully handle the unbalanced data, otherwise it may hurt the final performance. Since our usual practice is always to start with an individual SVM model training by using common imposter background dataset with the assumption that imposter may share a common subspace. This is a natural assumption but unbalanced data may construct an over-fitting hyper-plane for enrollment speaker.

Based on the above reasoning, we propose to take all the enrollment speakers as the instance of positive examples and all the imposter background dataset as the instance of negative examples to train a single universal SVM; all the support vectors from the imposter side will then be included into a new dataset, which can be called Universal Background Support (UBS) imposter dataset (Fig.2). Since the UBS imposter dataset is derived by pooling the information from all enrollment speakers, it is thus expected to span a more realistic imposter space and prevent over-fitting (Fig.2a) issue, which plagues many development stages.

Once the UBS imposter dataset is ready, it can be used to replace the original entire imposter background dataset to train the individual enrollment speaker SVM model. This is a simple method without the requirement of any complex configuration parameter tuning. It extracts the imposter data information from a "general" perspective.



(a) (b)

Figure 2: *Universal Background Support imposter dataset selection illustration. (a) Traditional SVM modeling for one-enrollment vs. all imposter;(b) Proposed UBS-based SVM modeling, all-enrollment vs. all imposter. "X" and "+" stands for negative example and positive example, respectively. The circled "X" and "+" stands for the support vector for negative example and positive example, respectively. All the circled "X" in (b) consists of UBS imposter dataset.*

Since the UBS imposter dataset selection method is an approach to find the more general imposter data, it may suffer in neglecting the "specific" imposter information for an enrollment speaker SVM model. If there is a way that can integrate more target-related specific imposter, then this more comprehensive imposter selection method is expected to contribute a more discriminative verification system. The reasoning for this comprehensive method is like the procedure followed in the traditional "UBM (universal background model)" and "adaptation", which is a classic solution in previous speaker verification/identification field to cope with the sparse training data issue. In the construction of SVM, each support vector in the model is attached a weighting coefficient, which indicates how much influence a given support vector has on the positioning of the hyperplane. So every group of imposter support vectors in an individual SVM model can be thought of as a possible imposter "adaptation" dataset source. To be specific, for the "UBM" step, the proposed method will be used to prepare the initial background imposter dataset and the support vector samples will be utilized to do the SVM model "adaptation".

To summarize, the proposed background data selection method can be implemented in the following four steps:

1. Using *all* the enrollment speakers data (noted as set ***E***) as positive examples and *all* the imposter speakers data (noted as set ***I***) as negative examples to train a universal SVM (noted as *a-SVM* );
2. Only using the negative support vectors of the *a-SVM* and the data of $j^{th}$ enrollment speaker to train the individual model for $j^{th}$ enrollment speaker (notes as $u\text{-}SVM_j$ ), where $1 \leqq j \leq S$, $S$ is the total number of enrollment speakers;
3. Using set ***I*** *as negative samples* and the data of the $j^{th}$ enrollment speaker to train the individual model for $j^{th}$ enrollment speaker (noted as $i\text{-}SVM_j$ , and there are $S$ *i-SVM*-type models in total);
4. Pooling the negative support vector from the $u\text{-}SVM_j$ and top $N$ most frequent negative support vectors from all the $S$ *i-SVMs* to form a new negative support vectors set (noted as set ***P***) , using the set ***P*** and the data of the $j^{th}$ enrollment speaker to train the individual model for the $j^{th}$ enrollment speaker (noted as $s\text{-}SVM_j$ ), which can be called specific or adapted SVM.

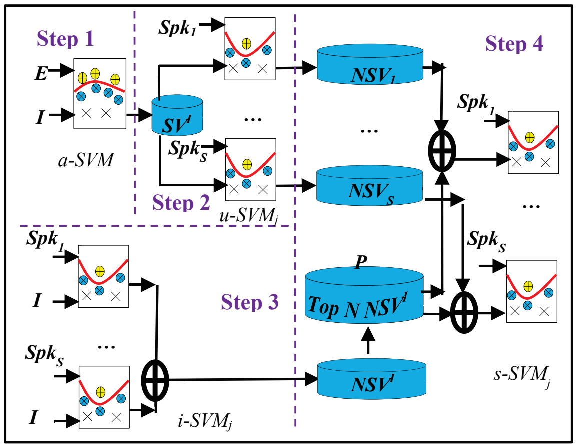


Figure 3: *Universal Background Support –step and Specific Step block diagram. NSV=Negative Support Vector; Spk=speaker; $NSV^A$ is the pooling of all the NSV from-SVM-type models, i.e,. $i\text{-}SVM_j$, where $1 \leq j \leq S$. Each step is isolated with purple dash line.*

Step 1 and 2 can be re-termed as *UBS-step* and Step 3 and 4 can be re-termed as *Specific-step* (or adaptation step). The whole scheme is also illustrated in Fig. 3.

# 4. Setup

## 4.1. System Setup

For parameterization, a 60-dimensional feature (19 MFCC with log energy +Δ+ΔΔ) using a 25ms analysis window with 10ms shift, filtered by feature warping using a 3-s sliding window is employed. The system also employs Factor Analysis, followed by Linear Discriminative Analysis (LDA) and Within Class Covariance Normalization (WCCN) for the SVM system [2]. SVM with cosine kernel is employed as the verification system here. *SVMlight* toolkit is used in our experiment [9]. Next, the NIST 2004, 2005, 2006 enrollment data are used to train the gender-dependent UBM 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 04, 05, and 06 male enrollment data with 5 or more recording sessions per speaker. A total of 400 factors were used. The LDA matrix is trained on the same data as the total variability matrix. In our experiments, the dimension of the LDA matrix is set to 140. Finally, the within class covariance matrix was trained using NIST04, and NIST05 data, and a cosine kernel was used in order to build the SVM systems. The system block diagram is shown in Fig. 1

To make a comprehensive evaluation, the performance of the proposed method will be compared against the traditional method (without using any imposter background selection) and the popular method (support vector frequency selection). To further explore the potential benefit of the proposed method, we also consider a comprehensive way that can integrate both "general" and "specific" imposter data information to build a more discriminative verification system.

## 4.2. Evaluation Dataset

All abovementioned algorithms are evaluated on the 5min-5min telephone-telephone core-condition of the NIST 2008 and 2010 speaker recognition evaluation (SRE) corpora. The evaluation dataset is limited to male speakers.

## 4.3. Background Dataset & Score Normalization Set

The background dataset consists of NIST04 and NIST05 with a total of 2718 utterances. Each utterance is parameterized and used as negative example. The *SVFreq* method is utilized to rank all the negative examples.

To make a clear presentation of the corpora structure, Tab. 1 summarizes all corpora that are used to estimate the UBM, JFA hyper-parameters, total variability matrix (T), LDA, NAP (nuisance attribute projection), WCCN, SVM model.

# 5. Results and Analysis

The system performance is measured in Equal Error Rate (EER) and the detection cost function (DCF), which follows the definition in NIST10 protocol [10].

## 5.1. Parallel comparison

Firstly, to better understand the reasoning behind the popular method (support vector selection method), we will apply it on NIST08 (development corpus) and NIST10 (evaluation corpora). The enrollment model of $j^{th}$ speaker is constructed by using two kinds of samples: positive samples (only one in the case of NIST08 and NIST10) and negative samples (varied size). The negative examples are actually speaker data from NIST04 and 05 (Tab. 1), which are ranked in descending order according to support vector frequency defined in Eq. 1. To do the imposter data selection, the size of ranked background dataset varied from 100 to 2000 with the step 100 in order to prevent the data selection from being reduced to be tedious searching. In another word, 100 means the top 100 most often used support vector from all individual SVM-based speaker enrollment model. The performance of varied size of background data on the two corpora is summarized in Fig. 4. Although, the SVFreq method can find a good operation point in the development corpus (EER=5.13% on NIST08) when the data size equals 500, its performance on the evaluation corpus of NIST10 is very poor. This method fails to offer consistent performance after time consuming parameter tuning on the development data. As a popular practice, in the system construction of both NIST08 and NIST10, all the background data come from same source, NIST04 and NIST05, but some variation exists between the data of NIST08 and NIST10. This can partially explain the inconsistence. To use the same background source data is the premise of the generalization of the method of SVFreq, so, this is a built-in shortcoming.

Table 1. *Corpora used to estimate the system components. ( Note: "X" means that data from this corpus was used)*

| Corpora | Switch-board | NIST 04 | NIST 05 | NIST 06 | NIST 08 | NIST 10 |
|---|---|---|---|---|---|---|
| **UBM** | X | X | X | X | | |
| **JFA Matrix V** | X | | X | X | | |
| **JFA Matrix D** | | X | | | | |
| **JFA Matrix U** | | X | X | X | | |
| **T** | X | X | X | X | | |
| **LDA** | X | X | X | X | | |
| **NAP** | | X | X | X | | |
| **WCCN** | | X | X | X | | |
| **SVM-TNorm** | | X | | | | |
| **SVM-imposter** | | X | X | | | |
| **Development** | | | | | X | |
| **Evaluation** | | | | | | X |

Secondly, we summarize in Tab. 2 the performance result of the proposed method and popular method, along with the performance without any background selection method. From

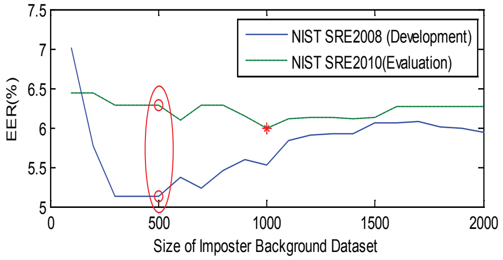


Figure 4: *Performance of varied background dataset size on development sets NIST08 and evaluation sets NIST10.*

Table 2. *Performance comparison of the proposed UBS method with baseline and SVFreq method.*

| Method | Development | Evaluation | EER | MinDCF |
|---|---|---|---|---|
| **Baseline** | / | 2008 | 6.01 | 0.602 |
| **Baseline** | / | 2010 | 6.14 | 0.615 |
| **SVFreq** | 2008 | 2008 | 5.13 | 0.513 |
| **SVFreq** | 2008 | 2010 | 6.28 | 0.628 |
| **SVFreq** | 2010 | 2010 | 6.00 | 0.610 |
| **UBS** | / | 2008 | 5.52 | 0.552 |
| **UBS** | / | 2010 | 5.69 | 0.570 |

Table 3. *Performance of comprehensive system: UBS+Specific*

| Selection Method | NIST SRE2008 | | NIST SRE2010 | |
|---|---|---|---|---|
| | **EER** | **MinDCF** | **EER** | **MinDCF** |
| **Baseline** | 6.01% | 0.602 | 6.14% | 0.615 |
| **UBS+Specific** | 4.90% | 0.490 | 5.67% | 0.567 |

Tab. 2, we can immediately notice that, unlike the method of *SVFreq*, the proposed method does not need tune any parameter therefore mast faster than the popular method. In addition, we can see the proposed method gives more consistent performance in both NIST08 and 10. It gives best performance in NIST10. This can be contributed to its better leveraging of the limited data in a more balanced way. The result from the proposed method in NIST08 is worse than the popular method. This can be due to the fact of mixed language: NIST08 contains mixed language, the imposter dataset (NIST04,05) also contains mixed language but NIST10 only contains English. This means the imposter data is more knowledgeable in NIST08 thus better performance is desirable. On the other hand, this performance difference suggests that the SVFreq method may be over-fitting on the development set and thus demonstrate weaker generalization capability in the less matched condition. The proposed method also needs no attempt on the tedious parameter searching over the development set. Actually, there are 2718 imposter examples in the initial entire imposter background dataset. So the total possibility will be astronomic (the factorial of 2718). The huge possibility provides a sophisticated structure which can approach any imposter space. At the same time, the astronomic possibility renders it impossible to do the exhaustive searching. A balance and well-informed imposter data selection method will be necessary.

### 5.2. Serial combination-Comprehensive method

The dataset derived from the proposed UBS method refer to the general information of all enrollment speakers. As stated in Sec.3, the "specific" information from an individual enrollment speaker-related imposter can adapt this "general" model to be more discriminative. Tab. 3 summarizes results.

Tab. 3, with Tab. 2, shows that the comprehensive system with both general and specific information is further improved in both verification tasks. Again, the extra benefit comes at the cost of expensive searching. It is noted that improvement of UBS+Specific method on SRE08 far outperforms that of SRE10, which may be explained by the fact that although they share many similar imposter data, they handle different enrollment data and matched condition favors NIST08 again.

## 6. Conclusions

This study has proposed a novel universal background support imposter cohort selection method within the SVM-based speaker verification system. Selection of an informative background dataset is crucial in the construction of a state-of-the-art SVM-based speaker verification system. Previous studies generally derive the optimal number of imposter examples from the development data and applied this to evaluation data; our evaluation shows this cannot give optimum results for unseen data (actually much worse performance than without using any method; Refer to Tab. 2). In the proposed method, a universal background dataset was derived to embed the imposter knowledge in a more balanced way. Next, the derived dataset was taken as the imposter sets in the SVM modeling process for each enrollment speaker. Compared to the popular support-vector frequency based method, the proposed method does not need parameter searching and also offers a 5.2% relative EER improvement on the NIST10 evaluation corpus. By employing UBM-Adaptation-like method, more target-specific imposter information can be embedded to build a well-informed model, which can be achieved at the cost of expensive computation.

The proposed method refers to the whole enrollment data information, and thus a more balanced and more general imposter dataset can be expected to help build the target SVM model to avoid potential over-fitting or under-fitting in the SVM model training stage. Since the enrollment data is always limited for a specific speaker, many studies focus on making good use of background imposter dataset. In addition, exploring the information from available ***enrollment*** data is not a trivial issue. In fact, in many cases, it is neglected. This study represents a preliminary exploration in this direction.

## 7. References

[1] V. Wan and W. Campbell, "Support vector machines for speaker verification and identification," in IEEE Workshop Neural Networks for Signal Processing, 2000, vol. 2, pp. 775-784

[2] S. S. Kajarekar and A. Stolcke, "NAP and WCCN: Comparison of approaches using MLLR-SVM speaker verification system," in Proc. IEEE ICASSP, 2007, pp. 249–252.

[3] S. S. Kajarekar, "Phone-based cepstral polynomial SVM system for speaker recognition," in Proc. Interspeech, 2008.

[4] M. McLaren, B. Baker, R. Vogt, and S. Sridharan, "Improved SVM speaker verification through data-driven background dataset collection," ICASSP-2009, pp. 4041- 4044, 2009.

[5] M. McLaren, B. Baker, R. Vogt, and S. Sridharan, "Exploiting multiple feature sets in data-driven impostor dataset selection for speaker verification," ICASSP-2010, pp. 4434-4437, 2010.

[6] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121-167, 1998.

[7] N. Brummer etc. "ABC system description for NIST SRE 2010". SRE Workshop 2010

[8] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front-end Factor Analysis for Speaker Verification," IEEE Transaction on Audio, Speech and Language Processing, 2010.

[9] T. Joachims, "SVMLight: Support Vector Machine," SVM-Light Support Vector Machine http://svmlight.joachims.org/, University of Dortmund, 1999.

[10] http://www.itl.nist.gov/iad/mig/tests/spk/2010/NIST_SRE10_early_registration.pdf