Queensland University of Technology
Brisbane Australia

This may be the author's version of a work that was submitted/accepted for publication in the following source:

Mclaren, Mitchell, Baker, Brendan, Vogt, Robert, & Sridharan, Sridha
(2010)
Exploiting multiple feature sets in data-driven impostor dataset selection for speaker verification.
In Douglas, S (Ed.) *Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing.*
Institute of Electrical and Electronics Engineers Inc., United States, pp. 4434-4437.

This file was downloaded from: https://eprints.qut.edu.au/32297/

**© Copyright 2010 IEEE**

©2010 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

**Notice**: *Please note that this document may not be the Version of Record (i.e. published version) of the work. Author manuscript versions (as Submitted for peer review or as Accepted for publication after peer review) can be identified by an absence of publisher branding and/or typeset appearance. If there is any doubt, please refer to the published source.*

*https://doi.org/10.1109/ICASSP.2010.5495620*

QUT Digital Repository:
http://eprints.qut.edu.au/

This is the author version published as:

McLaren, Mitchell L. and Baker, Brendan J. and Vogt, Robert J. and Sridharan, Sridha (2010) ***Exploiting multiple feature sets in data-driven impostor dataset selection for speaker verification.*** In: Proceedings of 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2010), 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas.

Copyright 2010 IEEE

# EXPLOITING MULTIPLE FEATURE SETS IN DATA-DRIVEN IMPOSTOR DATASET SELECTION FOR SPEAKER VERIFICATION

Mitchell McLaren, Brendan Baker, Robbie Vogt, Sridha Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia
{m.mclaren, bj.baker, r.vogt, s.sridharan}@qut.edu.au

## ABSTRACT

This study assesses the recently proposed data-driven background dataset refinement technique for speaker verification using alternate SVM feature sets to the GMM supervector features for which it was originally designed. The performance improvements brought about in each trialled SVM configuration demonstrate the versatility of background dataset refinement. This work also extends on the originally proposed technique to exploit support vector coefficients as an impostor suitability metric in the data-driven selection process. Using support vector coefficients improved the performance of the refined datasets in the evaluation of unseen data. Further, attempts are made to exploit the differences in impostor example suitability measures from varying features spaces to provide added robustness.

***Index Terms*—** speaker recognition, data selection, impostors, support vector machine

## 1. INTRODUCTION

Recent studies have highlighted the importance of selecting suitable impostor examples to form the background dataset in SVM-based speaker verification [1, 2, 3]. The selection of impostor datasets is often performed based on heuristics followed by a number of development evaluations. However, the major shortcoming of heuristic-based approach is that selection is not performed on a per-observation basis resulting in the sub-optimal dataset.

The proposal of data-driven impostor dataset refinement [2] addressed the shortcoming of the heuristic-based approach through the automated selection of individual impostor examples for the SVM background. This technique selects as a *refined* dataset the most appropriate subset of examples from a large and diverse candidate dataset. The suitability of impostor examples is measured using a metric based on the frequency of selection as a support vector for a set of development models.

While background dataset refinement has successfully been applied to the selection of the SVM background [2], SVM T-norm cohort [3] and both T- and Z-norm cohorts for GMM-based speaker verification [4], the refinement process has focussed solely on the use of GMM mean supervectors as features. The versatility of dataset refinement brings into question how effective the technique can be when applied to alternate SVM-based feature sets.

This study applies dataset refinement to a range of SVM-based feature sets in order to observe whether refinement can bring about benefits to kernels other than the GMM mean supervector kernel [5]. These feature sets include generalised linear discriminate sequences (GLDS) [6], phonetic N-grams [7] and maximum-likelihood linear regression (MLLR) transforms [8]. Further, the differences in candidate impostor ranking between several SVM configurations are exploited in an attempt to form a robust impostor suitability metric for the refinement of the candidate dataset in an unseen classifier.

This work also proposes a robust impostor suitability metric based on cumulative support vector coefficients. In contrast to the existing metric, support vector *frequency* [2], the new metric exploits the information contained in the weights or coefficients assigned to support vectors so as to provide increased resolution and robustness in the ranking of the candidate examples.

Section 2 describes the data-driven dataset refinement technique for impostor dataset selection. Section 3 defines each of the SVM configurations used in this work. Section 4 details the experimental protocol with results and discussions presented in Section 5.

## 2. DATA-DRIVEN IMPOSTOR DATASET SELECTION

This section describes the data-driven approach to impostor dataset selection recently proposed in [2]. An extension to the originally developed impostor suitability metric is also detailed in this section.

### 2.1. Support Vector Derived Impostor Suitability Metrics

The support vector machine is a discriminative classifier trained to separate classes in a high-dimensional kernel space by positioning a separating hyperplane in this space such that the maximum margin between classes is found [9]. The position of this hyperplane for a client SVM is defined by a set of informative training examples termed *support vectors* which are determined through the optimisation of the SVM objective function. The data-driven selection technique [2] is based on the assertion that the most informative background examples are regularly selected as support vectors during SVM training.

#### *2.1.1. Support Vector Frequency*

The *support vector frequency* [2] was originally proposed as a measure of a candidate example's relative importance in the background dataset. The support vector frequency of an example was defined as the number of times that it is selected as a support vector while training a set of SVMs on a development dataset. While this metric was found to be reliable in a number of scenarios [2, 3, 4], it does not fully exploit the information available from the support vector selection process, and instead, compresses the information given by the support vector coefficients.

This research was supported by the Australian Research Council (ARC) Discovery Grant Project ID: DP0877835.

#### 2.1.2. *Cumulative Support Vector Coefficients*

Further information can be exploited from the support vectors of a trained SVM than the fact that they were selected. Such information is held by the weight or coefficient assigned to each support vector. The support vector coefficient indicates how much influence a given support vector had on the positioning of the hyperplane. For example, the hyperplane normal of a trained SVM is given by $\boldsymbol{\omega} = \sum_i \alpha_i y_i \boldsymbol{x}_i$ where $\boldsymbol{x}_i$ is the $i^{\text{th}}$ training example with class label $y_i \in \{-1, 1\}$ and $\alpha_i$ is the coefficient assigned to the example. Those examples with a coefficient $\alpha_i > 0$ are defined as support vectors, while examples with $\alpha_i = 0$ had no impact on the final position of the hyperplane. An example that is allocated a relatively large coefficient has a corresponding impact on the hyperplane.

Information from support vector coefficients can be exploited to form an impostor suitability metric — cumulative support vector coefficients. This metric defines the suitability of a candidate impostor example as the sum of it's allocated coefficients during the training of a development set of client SVMs. Specifically, the cumulative support vector coefficients of background example $j$ can be formulated as,

$$\text{SVCoef}_j = \sum_{k=1}^{K} \alpha_j^k \tag{1}$$

from a total of $K$ client hyperplanes trained using a candidate background dataset of $J$ impostor examples (ie., $j \in y_i = -1$).

In contrast to (1), the support vector frequency implies a ceiling function on $\alpha_j^k$ to achieve a 'count' statistic. The cumulative support vector frequency, therefore, has the advantage of increased resolution over the support vector frequency for the ranking of candidate impostor examples.

### 2.2. Background Dataset Refinement

The process of data-driven dataset refinement was previously detailed in [2]. This technique refines a diverse set of vectors $B$, compiled from a number of available resources into a suitable impostor dataset using a set of development client vectors $S$. Ranking of the candidate dataset $B$ involves, firstly, training a set client SVMs for each observation in $S$ using $B$ as background examples. The impostor suitability of each example in $B$ is then be calculated using (1) after which the entire dataset can be ranked. From this ranked dataset, a *refined* dataset $R_N$ is formed as the $N$ highest-ranking examples of $B$.

## 3. SVM CONFIGURATIONS

The SVM configurations used in this study employ both cepstral-based and high-level feature sets. The application of dataset refinement to these different feature sets is expected to demonstrate the versatility of the technique.

### 3.1. GMM Mean Supervector SVM

The GMM mean supervector SVM system used in this study was previously described in [2] and has been the system utilised in all previous studies on dataset refinement [2, 3, 4].

GMM supervectors were produced through mean-only MAP adaptation using 24-dimensional, feature-warped MFCC features with appended delta coefficients. An adaptation relevance factor of $\tau = 8$ and 512-component models were used throughout. SVM training and classification was performed using the GMM mean supervector kernel [5] with GMM supervectors of 12288 dimensions. This system is denoted *GMM-Svec* in this work.

### 3.2. Generalised Linear Discriminant Sequence Features

A derivative of the generalised linear discriminate sequence (GLDS) system proposed by Campbell, *et al.* [6] is adopted in the following experiments to observe whether dataset refinement is able to provide benefits to an alternate sequence kernel to the GMM mean supervector SVM. The GLDS features are derived from a generalised discriminant function to represent a speech segments as a vector of scalar functions.

The GLDS implementation in the following study produces a vector, $\boldsymbol{b}(\boldsymbol{x})$, as a set of polynomial basis terms for each input feature vector. An example of a 2nd degree polynomial function for an input feature vector of two dimensions, $\boldsymbol{x} = [\boldsymbol{x}_1 \quad \boldsymbol{x}_2]^t$, is given by,

$$\boldsymbol{b}(\boldsymbol{x}) = \begin{bmatrix} 1 & \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_1^2 & \boldsymbol{x}_1\boldsymbol{x}_2 & \boldsymbol{x}_2^2 \end{bmatrix}^t. \tag{2}$$

In this work, MFCC feature vectors of 24 dimensions are utilised to produce the 4th degree ($N = 4$) polynomial basis terms. This results in SVM input vectors of 20475 dimensions. Given (2), a set of MFCC feature vectors $\boldsymbol{X} = \{\boldsymbol{x_0}, \boldsymbol{x_1}, \cdots, \boldsymbol{x_n}\}$ extracted from an utterance can then be conveniently represented as a finite vector for SVM-based training and classification using $\hat{\boldsymbol{b}}(\boldsymbol{X}) = \frac{1}{n}\sum_i^n \boldsymbol{b}(\boldsymbol{x_i})$. Features of the form $\hat{\boldsymbol{b}}$ are utilised in this study along with the non-parametric rank normalisation kernel [10] as opposed to the GLDS kernel structure originally implemented in [6].

### 3.3. English Phonetic Lattice N-Gram Features

The phonetic lattice 'bag-of-N-grams' classifier [7] aims to capture speaker idiosyncrasies from a speech signal by modelling the probabilities of a N-phone sequences from the phonetic transcripts of speakers utterances. The system implemented in the following experiments used phonetic transcripts produced by gender-dependent, English phone recogniser developed within the QUT SAIVT laboratory and described in [11]. The system is capable of recognising a total of 43 phonetic labels. Similar to the work in [7], phonetic lattices were used to represent the transcriptions rather than 1-best transcriptions to allow for increased information regarding phonetic probabilities at the cost of additional computation.

Feature extraction involved concatenating the expected frequencies of unigrams, bigrams and trigrams in each utterance into a feature vector of 11893 dimensions. Only the 10,000 most frequently occurring trigrams (determined on a held-out dataset) were included. The N-gram frequencies were weighted according to their posterior probability of occurrence in the recognition lattice.

### 3.4. MLLR Transforms as Features

Maximum-likelihood linear regression (MLLR) transforms [12] are trained using cepstral-based features, however, they also normalise the underlying statistics for the choice of phones and word used in a speech signal. The training of MLLR transforms is integrated into many phone recogniser systems to model the differences between the world language model and the characteristics exhibited by the speaker. Consequently, the use of these transforms as features is able to provide a high degree of speaker-discriminative information when applied to SVM-based classification [8].

|  | SRE'06 |  | SRE'08 |  |
|---|---|---|---|---|
| **Background** | DCF | EER | DCF | EER |
| Complete | .0152 | 3.20% | .0185 | 4.07% |
| Refined (SVFreq) | **.0120** | **2.49%** | .0168 | 4.17% |
| Refined (SVCoef) | .0124 | 2.55% | **.0166** | **3.83%** |

**Table 1**. Comparison of support vector frequency and cumulative support vector coefficients as the impostor suitability metrics for dataset refinement on the GMM-Svec configuration.

The MLLR system implemented in this study uses the same Fisher trained English phone recogniser as used in the N-gram configuration detailed in Section 3.3. The male and female English phone recogniser HMM models served as reference models for computing speaker-dependent MLLR transforms. Using the alignments produced by the phonetic decoder, a five class regression tree was used (4 data driven classes + silence) to obtain a set of MLLR transforms for each training segment. The transform components for each class (excluding the silence class) were concatenated to form a single feature vector for each conversation side. Utterances were represented by 12480 dimensions by concatenating the affine transforms from both male and female trained models to form a gender-independent feature set.

## 4. EXPERIMENTAL CONFIGURATION

Training and classification of the GLDS, N-Gram and MLLR features sets was performed using a linear, non-parametric rank normalisation kernel [10]. This technique replaces each element of an input vector with its corresponding element rank value in a large set of held out vectors, after which rank values are normalised to the range [0, 1]. Nuisance attribute projection (NAP) [13] was incorporated in all systems and was performed subsequent to rank normalisation (where applicable) with 50 session dimensions being removed.

Large gender-dependent impostor datasets $B$ were collected from NIST 2004 and NIST 2005 databases and a random selection of 2000 utterances from each of Fisher and Switchboard 2 corpora giving a total of 6430 male and 7739 female observations. The number of examples from each of these data sources was similar to those used in [2]. For this study, these datasets consisted only of telephony data. Conversations were spoken in a range of languages with the majority in English.

The gender-dependent development client datasets $S$ used to calculate the impostor suitability metrics were compiled from the English training and testing utterances in the 1conv4w condition of the NIST 2006 SRE. Performance was evaluated on both SRE'06 and SRE'08. All NIST 2008 results were derived from condition 7 as specified in the official evaluation protocol which restricts trials to English spoken telephony data.

## 5. RESULTS

### 5.1. Comparing Impostor Suitability Metrics

A comparison of the support vector frequency (SVFreq) and cumulative support vector coefficient (SVCoef) metrics for impostor suitability, as defined in Section 2.1, was initially performed in this study using the GMM-Svec system. The candidate dataset $B$ was ranked using each metric from which refined datasets were selected to minimise EER on the SRE'06. Results obtained on both the SRE'06 and SRE'08 when evaluated using the refined datasets are detailed in Table 1. Results indicate that, although the frequency metric produced a superior refined dataset for the development conditions, the metric based on cumulative coefficients allowed the refined dataset to generalise better to the evaluation of unseen data.

|  |  | SRE'06 |  | SRE'08 |  |
|---|---|---|---|---|---|
| **System** | **Background** | DCF | EER | DCF | EER |
| GMM-Svec | Complete | .0152 | 3.20% | .0185 | 4.07% |
|  | Refined | **.0124** | **2.55%** | **.0166** | **3.83%** |
| GLDS | Complete | .0253 | 5.30% | .0303 | **6.51%** |
|  | Refined | **.0223** | **4.66%** | **.0289** | 6.76% |
| N-Gram | Complete | .0330 | 7.10% | .0414 | 9.78% |
|  | Refined | **.0290** | **6.39%** | **.0394** | **9.18%** |
| MLLR | Complete | .0311 | 6.99% | .0389 | 9.28% |
|  | Refined | **.0300** | **6.72%** | **.0380** | **9.04%** |

**Table 2**. T-normalised minimum DCF and EER obtained from 1-sided, English-only SRE'06 and SRE'08 when using the complete and refined background datasets in different SVM configurations.

These results demonstrate that cumulative support vector coefficients allow dataset refinement to better achieve it's intended objective — to select a suitable dataset that generalises well to the evaluation of unseen data. This impostor suitability metric will, therefore, be utilised for the remaining experiments in this study.

### 5.2. NIST Corpora Evaluations

Experiments in this section investigate the benefits that dataset refinement offers to each of the SVM-based configurations detailed in Section 3. The candidate dataset $B$ was individually refined for each SVM configuration with the SRE'06 corpus serving as development data. The cumulative support vector coefficient (SVCoef) was used as the impostor suitability metric. In all cases, the size of the refined dataset was selected to minimise the SRE'06 EER.

Table 2 details the performance statistics obtained when evaluating the NIST SRE'06 and SRE'08 copora with the refined and full candidate datasets in each SVM system. The refinement of the candidate dataset consistently brought about superior performance statistics from each SVM system in the development evaluations. The most significant improvements from dataset refinement were observed in the GMM-Svec system with a relative reduction of 18% in minimum DCF and 20% in EER over the candidate impostor dataset.

The SRE'08 results in Table 2 demonstrate that the refined datasets generalised well to unseen data such that performance improvements were observed relative to the candidate dataset in the majority of cases. Relative improvements of up to 10% were observed in the performance metrics from the refined datasets over the use of the candidate dataset along with a substantial reduction in dataset size of 80-85%. One inconsistency in the SRE'08 results was the loss in EER of the GLDS system when using the refined dataset. The reason for this drop in performance is expected to have arisen due to the relatively large kernel space of the GLDS configuration compared to the alternate SVM systems. The availability of a large number of dimensions may have allowed the refinement process to produce a dataset that was over-refined towards the development data, thereby reducing it's potential performance on unseen data. Further investigations are required to determine the extent that the size of the kernel space effects the refinement process.

Overall, these results demonstrate that background dataset refinement was able to bring about performance improvements on the differing feature sets of each SVM configuration trialled.

|  | Appears in # Sets |  |  |  |
|---|---|---|---|---|
|  | $\geq 1$ | $\geq 2$ | $\geq 3$ | ALL |
| No. Examples | 3588 | 2303 | 1400 | 709 |

**Table 3**. Number of impostor examples that appear in the top 2000 from each SVM feature set. Results for male subset.

### 5.3. Exploiting Inter-Feature Impostor Suitability Metrics

The following experiments investigate, firstly, the similarity of the ranked candidate datasets from the four SVM configurations and, secondly, whether the cumulative coefficient metrics obtained in several kernel spaces can be combined to form a robust impostor suitability metric for use in an unseen classifier.

In order to observe the extent of correlation between the ranking of the candidate dataset in each SVM configuration, the 2000 highest-ranking candidate examples from each system were analysed. Table 3 shows the overlap and how the candidate examples were distributed across the datasets. It can be seen that in the four sets, 3588 unique examples were observed, 709 of which were common to the top 2000 of all four systems. Whilst there is significant overlap between the sets, a total of 1285 examples (3558-2303=1285) appeared in only one of the four datasets. Thus, on average, around 16% of examples from each dataset were unique to that system.

Given the different composition of ranked candidate datasets from each SVM configuration, of interest is whether these differences can be exploited in order to form a more general ranking metric to be utilised in the refinement process. To evaluate the potential benefits gained through combination of impostor suitability measures, a set of experiments looked at combining measures from the GMM-Svec, MLLR and N-gram systems in order to determine an appropriate background set for a held-out classifier, in this case, the GLDS system. The previous experiments using the GLDS features (Section 5.2) resulted in limited gains in the evaluation of unseen data through system-dependent refinement. By combining several metrics from alternate feature sets, it was anticipated that a refined dataset for the GLDS system could be more robustly selected.

The approaches to metric combination trialled include a minimum, maximum and averaging function of the metrics. In these experiments, the minimum metric is analogous to the intersection of the datasets, while the maximum is similar to finding their union. Each of these approaches were used in the ranking of the candidate dataset from which refined datasets were selected to minimise the SRE'06 EER in the GLDS configuration (with possible differences in background sizes) before being evaluated on the SRE'08. Table 4 details the results from these evaluations. For comparison, performance from the refined dataset selected via system-dependent refinement with GLDS features is also presented in this table.

The SRE'06 results in Table 4 indicate that performance offered through each metric combination approach was comparable to the use of system-dependent refinement in the GLDS system. In contrast, the use of the maximum impostor suitability metric in the SRE'08 resulted in the best performance in the evaluation of the unseen dataset; superior even to system-dependent impostor suitability measures. These results suggest that the combination of cumulative support vector coefficients from multiple feature sets disjoint to those in the classifier used for refinement can aid in producing an appropriately refined dataset for the evaluation of unseen data.

## 6. CONCLUSION

This study investigated the applicability of dataset refinement to the selection of the background dataset for a range of SVM-based feature sets. These sets included GLDS features based on a polynomial expansion, English phonetic 'bag-of-N-Grams', MLLR transforms and the GMM mean supervector system. Development evaluations found that dataset refinement provided performance improvements in all SVM systems relative to the use of the un-refined dataset. The refined datasets selected on a system-dependent basis were also found to generalise well to the unseen data of the SRE'08.

| Metric | Bckgnd. | SRE'06 |  | SRE'08 |  |
|---|---|---|---|---|---|
| Choice | Size | DCF | EER | DCF | EER |
| GLDS | 1000 | .0223 | **4.66%** | .0289 | 6.76% |
| Minimum | 750 | **.0222** | 4.71% | .0300 | 6.83% |
| Average | 1250 | .0227 | 4.71% | .0290 | **6.59%** |
| Maximum | 1250 | .0225 | **4.66%** | **.0283** | 6.61% |

**Table 4**. GLDS performance obtained when combining the impostor suitability metrics from alternate systems for the refinement process.

An impostor suitability metric based on cumulative support vector coefficients allowed the refinement process to select a dataset that generalised better to unseen data than the use of the originally proposed frequency metric. Finally, the combination of suitability metrics from multiple classifiers provided added robustness for the ranking of a dataset prior to it's refinement in a held out classifier.

## 7. REFERENCES

[1] A. Stolcke, S. S. Kajarekar, L. Ferrer, and E. Shriberg, "Speaker recognition with session variability normalization based on MLLR adaptation transforms," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, pp. 1987–1998, 2007.

[2] M. McLaren, B. Baker, R. Vogt, and S. Sridharan, "Improved SVM speaker verification through data-driven background dataset selection," in *Proc. ICASSP*, 2009, pp. 4041–4044.

[3] M. McLaren, B. Baker, R. Vogt, and S. Sridharan, "Data-driven impostor selection for T-norm score normalisation and the background dataset in SVM-based speaker verification," in *Proc. ICB*, 2009, pp. 474–483.

[4] M. McLaren, R. Vogt, and S. Sridharan, "Improved GMM-based speaker verification using SVM-driven impostor dataset selection," *To be presented in Proc. Interspeech*, 2009, 2009.

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

[6] W.M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," *Proc. IEEE ICASSP*, vol. 1, pp. 161–164, 2002.

[7] A.O. Hatch, B. Peskin, and A. Stolcke, "Improved phonetic speaker recognition using lattice decoding," in *Proc. IEEE ICASSP*, 2005, vol. 1.

[8] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR Transforms as Features in Speaker Recognition," in *Proc. Eurospeech*, 2005, pp. 2425–2428.

[9] 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.

[10] A. Stolcke, S. Kajarekar, and L. Ferrer, "Nonparametric feature normalization for SVM-based speaker verification," in *Proc. IEEE ICASSP*, 2008, pp. 1577–1580.

[11] B.J. Baker, R. Vogt, M. McLaren, and S. Sridharan, "Scatter difference NAP for SVM speaker recognition," in *Proc. ICB*, 2009, pp. 464–473.

[12] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[13] A. Solomonoff, C. Quillen, and W.M. Campbell, "Channel compensation for SVM speaker recognition," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2004, pp. 57–62.