Data-Driven Impostor Selection for T-Norm Score Normalisation and the Background Dataset in SVM-Based Speaker Verification

Mitchell McLaren, Robbie Vogt, Brendan Baker, and Sridha Sridharan

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

Abstract. A data-driven background dataset refinement technique was recently proposed for SVM based speaker verification. This method selects a refined SVM background dataset from a set of candidate impostor examples after individually ranking examples by their relevance. This paper extends this technique to the refinement of the T-norm dataset for SVM-based speaker verification. The independent refinement of the background and T-norm datasets provides a means of investigating the sensitivity of SVM-based speaker verification performance to the selection of each of these datasets. Using refined datasets provided improvements of 13% in min. DCF and 9% in EER over the full set of impostor examples on the 2006 SRE corpus with the majority of these gains due to refinement of the T-norm dataset. Similar trends were observed for the unseen data of the NIST 2008 SRE.

1 Introduction

An issue commonly faced in the development of speaker verification systems is the selection of suitable datasets. Several recent studies have highlighted the importance of selecting appropriate training, development or impostor data for SVM-based speaker verification to match the evaluation conditions [1,2]. Two datasets fundamental to SVM-based classification that must be appropriately selected to maximise classification performance are the background dataset and the dataset used for test score normalisation (T-norm) [3].

The background dataset is a collection of negative or impostor observations used in the training of an SVM in which discrimination between the background dataset and the speaker examples is maximised [4]. Often, the number of impostor observations significantly outweighs that of speaker examples such that the SVM system relies heavily on the background observations to provide most of the observable discriminatory information. The background dataset must, therefore, consist of suitable impostor examples to ensure good classification performance.

Similar to the background dataset, the T-norm dataset is a collection of impostor examples from which a set of T-norm models are trained. T-normalisation uses these models during testing to estimate an impostor score distribution with

M. Tistarelli and M.S. Nixon (Eds.): ICB 2009, LNCS 5558, pp. 474–483, 2009.

which classification scores are normalised [3]. Consequently, the reliable estimation of impostor score distributions, and ultimately, the potential gains offered through T-norm are dependent on the appropriate selection of T-norm dataset.

Data-driven background dataset refinement [5] was a recently proposed technique to individually assess the suitability of each candidate impostor example, from a large and diverse dataset, for use in the background dataset. The support vector frequency of an example was used as a measure of its relative importance in the background dataset to rank the set of impostor examples. The top N examples from this ranked set were then used as a *refined* background dataset to provide improved classification performance. The system configuration in this study refined a single dataset to be used as both the background and the T-norm dataset. Of interest, however, is the way in which classification performance is affected when the size of the refined background and T-norm datasets are allowed to vary independently.

The following study investigates how sensitive SVM-based classification performance is to the selection of a suitable T-norm dataset compared to the selection of the background dataset. This is analysed by observing performance when independently varying the number of highest-ranking candidate impostor examples in the refined background and T-norm datasets. The ranking of candidate impostor examples is performed using the same refinement procedure outlined in [5].

The recently proposed data-driven background dataset refinement technique is presented in Section 2 followed by a discussion in Section 3 on T-norm score normalisation in SVM-based classification. Section 4 details the experimental protocol with results presented in Section 5.

2 Data-Driven Background Dataset Selection

Data-driven background dataset refinement [5] was recently shown to be an effective method for the selection of a highly informative background dataset from a set of candidate impostor examples such that it exhibited performance gains over the best heuristically-selected background from the same initial resources. The technique makes use of a development dataset to systematically drive the selection of the impostor dataset based on the relevance of each example in the background dataset. In this approach, the ranking of impostor observations is performed using a criterion that involves exploiting the information possessed by the support vectors of a trained SVM.

2.1 Support Vector Frequency

The support vector machine is a discriminative classifier trained to separate classes in a high-dimensional space. A kernel is used to project input vectors into this high-dimensional space where a separating hyperplane is positioned to maximise the margin between the classes [4]. The training of a speaker SVM results in the selection of a subset of both positive and negative examples from the

training dataset termed *support vectors* that are used to construct the separating hyperplane. Examples that were selected as support vectors in the SVM hold a common property of being the most difficult to classify, lying on, or within, the margin between classes. In contrast, those training examples that were not selected as support vectors provided no information in the training of the SVM.

The process of determining a subset of support vectors during SVM training can be considered a data selection process in which the most informative examples are chosen from the training dataset. In light of this, it can be stated that the impostor support vectors are the most informative set of background examples with respect to the client data.

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 of an example is defined as the number of times that it is selected as a support vector while training a set of SVMs on a development dataset.

2.2 Background Dataset Refinement

Given a diverse set of vectors B, compiled from a number of available resources, this dataset can be refined into a suitable background dataset using a set of *development* client vectors S. The speakers and vectors in the set S should be disjoint from those in B.

1. Using the full set of impostors B as the background dataset, train SVMs for each vector in the set of development client models S.
2. Calculate the support vector frequency of each impostor example in B as the total number of instances in which it was selected as a support vector for the development client models.
3. The refined impostor dataset R_N is chosen as the top N subset of B ranked by the support vector frequency ($R_N \subset B$).
4. For several values of N, use R_N in the evaluation of a development corpus to determine the optimal number of examples to be included in the refined background dataset.

It is important to note that the support vector frequencies are likely to be heavily dependent on the characteristics found in the development set S. For this reason, S should be selected based on the knowledge of the broad characteristics (such as gender, language and audio conditions) expected to exist in the corpus for which the impostor dataset is intended to be used.

3 Test Score Normalisation

Test score normalisation (T-norm) [3] is a technique used to counteract the statistical variations that occur in classification scores and was found to be an integral part of most speaker verification systems submitted to the recent NIST

speaker recognition evaluations (SRE) [6]. T-norm aims to normalise the score distributions of all test segments to a single scale so as to improve performance when applying a global threshold to test classification scores. This involves scoring each test utterance on a set of impostor models, trained from the T-norm dataset, producing an impostor score distribution with mean μ_I and standard deviation σ_I. The score s, obtained when comparing the test segment to a client model, is then normalised using,

$$\overline{s} = \frac{s - \mu_I}{\sigma_I} \tag{1}$$

The reliable estimation of the normalisation parameters μ_I and σ_I is dependent on the observable characteristics of the T-norm dataset. Consequently, the objective of normalising scores to a global scale will be more attainable if the selection of the T-norm dataset is tailored toward the expected evaluation conditions.

3.1 T-Norm Dataset Selection for SVM-Based Classification

The desired characteristics of the T-norm dataset closely match those wanted of a background dataset, in being a set of examples that appropriately represents the impostor population. Campbell et al. demonstrated this commonality of requirements by comparing the use of a single dataset for both the background and T-norm datasets to the use of disjoint datasets [7]. It was found that performing T-norm using the background dataset provided an improvement in performance over unnormalised scores, while the disjoint T-norm dataset resulted in somewhat degraded performance. It is unclear why this degradation occured, however, possible explanations include mismatch between the T-norm and evaluation conditions and also the limited size of the T-norm dataset.

The use of a single impostor dataset for the background and T-norm datasets was further explored in the recent study on data-driven background dataset refinement [5]. The evaluation of a development corpus demonstrated consistent performance gains as the large and diverse background dataset was more extensively refined. Corresponding gains were also observed in the evaluation of an unseen corpus.

The following study extends on the research in [5] by investigating the degree that classification performance depends on the suitable selection of the T-norm dataset compared to that of the background dataset. In contrast to the selection of a single dataset, *intersecting* datasets will be formed through the *independent* selection of the top N examples for the T-norm and background datasets from a single, ranked impostor dataset. In this way, the smaller dataset will be a subset of the other. Analysis of performance over a range of dataset sizes is expected to provide insight as to how sensitive SVM classification performance is to the selection of each of these datasets and whether the characteristics of the best refined background dataset are in fact similar those of the best T-norm dataset.

While this study will focus on the data-driven selection of *intersecting* background and T-norm datasets, the use of refined, *disjoint* datasets will also be evaluated for completeness.

4 Experimental Protocol

4.1 GMM-SVM System

SVM classification in the following experiments was based on GMM mean supervectors using the associated GMM mean supervector kernel [8]. The GMM system used in this study was based on 512-component models and was previously described in [9].

The SVM implementation uses the open source LIA_MISTRAL package [10] based on the libSVM library [11]. Nuisance attribute projection (NAP) [8] was employed to reduce session variation with the 50 dimensions of greatest session variation being removed from all supervectors.

4.2 Evaluation Datasets

Gender-dependent background datasets were collected from NIST 2004 and NIST 2005 databases and a random selection of 2000 utterances[1] from each of Fisher and Switchboard 2 corpora giving a total of 6444 male and 7766 female observations. The number of impostor examples from each of these data sources can be found in Table 1. The limited amount of data from the NIST 2005 corpus is due to the intentional exclusion of utterances from any speakers that also appear in the NIST 2006 corpus. For this study, these datasets consisted only of telephony data. Conversations were spoken in a range of languages with the majority in English. Large gender-dependent background datasets B were compiled from all available resources as listed in Table 1.

The gender-dependent development client dataset S used to calculate support vector frequencies was compiled from the training and testing utterances in the all-language, 1conv4w condition of the NIST 2006 SRE. Consisting of 1950 male and 2556 female client vectors, this provided a moderate degree of resolution in the support vector frequency statistic.

The NIST 2008 SRE corpus was used to observe how well the refined background and T-norm datasets generalised to unseen data. All NIST 2008 results were derived from condition 6 as specified in the official evaluation protocol [12] which includes trials spoken in all-languages while being restricted to telephony data, matching the conditions found in the development dataset S.

5 Results

5.1 Development Evaluations

Figure 1 depicts 3-D plots of the min. DCF and EER obtained on the NIST 2006 development corpus when using a range of refined T-norm and background dataset sizes selected as the highest-ranking impostor observations from the full dataset B. The darker peaks in the plots designate improved performance.

Figure 1(a) indicates that the min. DCF was more sensitive to the selection of a suitable T-norm dataset than the background dataset. This is evident in the higher

[1] Selected randomly due to memory limitations restricting the full dataset size.

Table 1. Number of impostor examples from each data source

Gender	Fisher	SWB2	NIST04	NIST05
Male	2000	2000	1901	543
Female	2000	2000	2651	1115

performance variation along the dimension corresponding to the size of the T-norm dataset compared to that of the background dataset. In contrast to the min. DCF, the EER plot in Figure 1(b) appears to exhibit less consistent dataset dependencies. As the background dataset is increasingly refined, however, the sensitivity of performance to the selection of the T-norm dataset appears to become clearer. Interestingly, the EER begins to degrade quicker than the min. DCF as the T-norm and background datasets are refined too extensively.

Maximum classification performance was found when using the top 1000 ranking impostor examples as the T-norm dataset and the subset of the top 250 observations as the background dataset. These datasets will be designated by the notation Bkg=B_{250} and T-norm=B_{1000}. Results from development evaluations using these refined datasets and the full impostor set B are detailed in Table 2. The use of both refined datasets provided a relative gain of 13% in min. DCF and 9% in EER over the full dataset B which is a statistically significant improvement at the 99% and 95% confidence level[2], respectively. With T-norm=B, the refined background dataset offered performance improvements over the full background, however, superior gains were observed from the refined T-norm dataset over the full T-norm set when Bck=B. These results demonstrate, firstly, that background dataset refinement can sucessfully be applied to the task of T-norm dataset selection, and secondly, that SVM-based classification is more dependent on the selection of a suitable T-norm dataset than the background dataset.

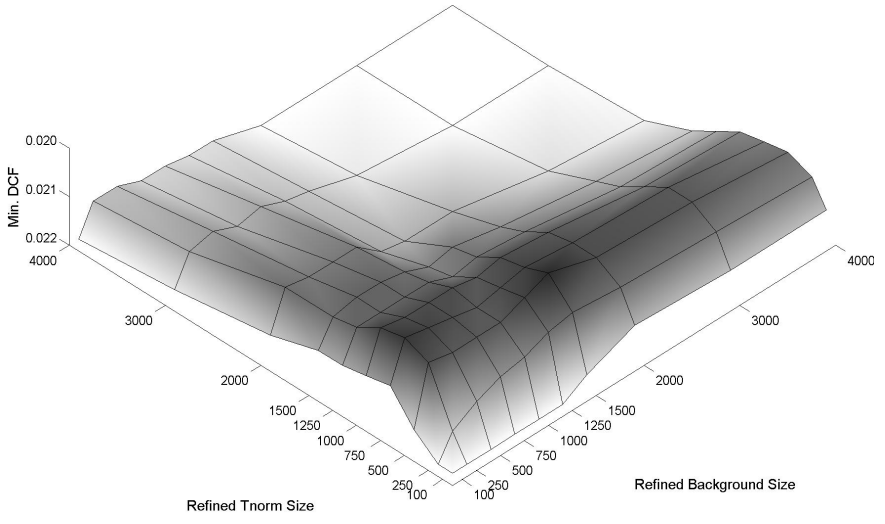
5.2 NIST 2008 Evaluations

The objective of this section was to determine whether the dataset-dependent performance trends, observed in the previous section (Section 5.1), were also reflected in the evaluation of the unseen data of the NIST 2008 SRE. Figure 2 depicts the 3-D plot of the min. DCF from these evaluations as the full set of impostor examples B was refined after being ranked using the NIST 2006 SRE corpus as development data.

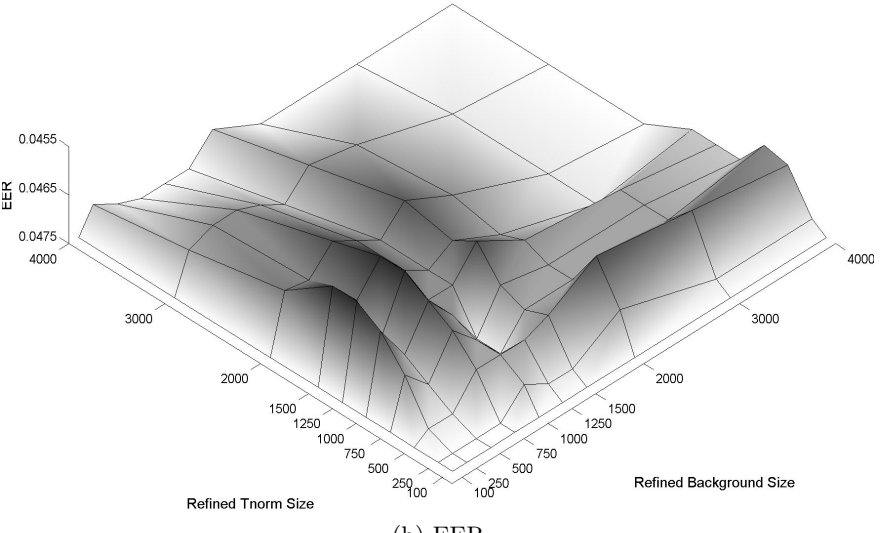
Figure 2 indicates that the NIST 2008 SRE min. DCF performance was more sensitive to the selection of a suitable T-norm dataset than that of background dataset selection, thus supporting the observations from development evaluations in Section 5.1. In contrast to the background dataset, the T-norm dataset appears to provide relatively high gains through increased refinement as observed by the darker regions of the plot.

Results from NIST 2008 SRE when using the datasets selected based on NIST 2006 development evaluations (Bkg=B_{250} and T-norm=B_{1000}) are detailed in Table 3 along with results obtained using the full impostor set. The use of the

[2] Based on the proposed method in [13] (independent case).



(a) Min. DCF



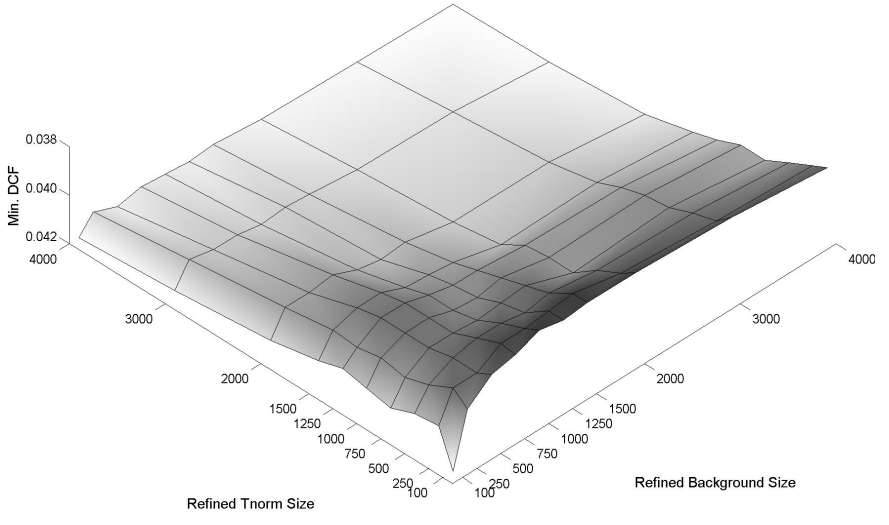
(b) EER

Fig. 1. Min DCF. and EER on NIST 2006 SRE when performing data-driven impostor selection of intersecting background and T-norm datasets

refined T-norm dataset offers substantial improvements over the full dataset while, surprisingly, the refined background dataset achieves comparable performance to the full dataset. These results demonstrate that the selection of a suitable T-norm dataset can have more impact on potential classification performance than the background dataset in the evaluation of an unseen corpus. When using both the refined datasets, a statistically significant improvement of 8% was observed in the min. DCF at the 99% confidence level[2] over the full datasets, however, no gain was found in EER.

Table 2. Performance on NIST 2006 SRE when using full dataset B and best refined intersecting T-norm and background datasets

Config. (Bck / T-norm)	Bck	T-norm	Min. DCF	EER
Full / Full	B	B	.0234	5.06%
Full / Refined	B	B_{1000}	.0215	4.70%
Refined / Full	B_{250}	B	.0223	4.84%
Refined / Refined	B_{250}	B_{1000}	.0203	4.59%



(a) Min. DCF

Fig. 2. Min DCF. on NIST 2008 SRE when using refined intersecting background and T-norm datasets ranked using NIST 2006 data

Table 3. Performance from NIST 2008 SRE using full and best refined T-norm and background datasets selected based on NIST 2006 evaluations

Config. (Bck / T-norm)	Bck	T-norm	Min. DCF	EER
Full / Full	B	B	.0435	8.34%
Full / Refined	B	B_{1000}	.0408	8.18%
Refined / Full	B_{250}	B	.0421	8.57%
Refined / Refined	B_{250}	B_{1000}	.0399	8.35%

5.3 Refinement of Disjoint Datasets

The refinement of a single dataset to form intersecting background and T-norm datasets provided a suitable means of investigating the dependence of SVM-based classification performance to their selection. The most common dataset configuration, however, involves the use of disjoint datasets as observed in recent NIST SRE submissions [6]. This section endeavours to determine firstly,

Table 4. Performance on NIST 2006 and NIST 2008 SRE using full and best refined *disjoint* T-norm and background datasets

Config. (Bck / T-norm)	Bck	T-norm	2006 SRE DCF	2006 SRE EER	2008 SRE DCF	2008 SRE EER
Full / Full	I	T	.0230	4.93%	.0432	8.34%
Full / Refined	I	T_{750}	.0214	4.73%	.0408	8.44%
Refined / Full	I_{250}	T	.0222	4.86%	.0428	8.37%
Refined / Refined	I_{250}	T_{750}	.0202	4.48%	.0394	8.30%

whether similar T-norm-dependence trends are observed when using disjoint datasets, and secondly, whether refined disjoint datasets provide performance improvements over refined intersecting datasets.

The full set of impostor examples B was divided to form the unrefined disjoint T-norm and background datasets T and I respectively, such that the speakers and vectors in these sets were separate. These disjoint subsets contained similar proportions of examples from each of the data sources listed in Table 1. Ranking of these sets was performed independently using the NIST 2006 corpus.

The refined disjoint datasets providing maximum performance in the NIST 2006 development evaluations were Bkg=I_{250} and T-norm=T_{750}. Results from trials on both NIST 2006 and NIST 2008 SRE using these datasets are detailed in Table 4. In the evaluation of both corpora, the refined T-norm dataset demonstrated substantial improvements over the full T-norm dataset. In contrast, the refined background provided somewhat improved results in the development evaluations, however, these benefits were only reflected in the NIST 2008 SRE when used in conjunction with the refined T-norm dataset. These results demonstrate that, even in the case of disjoint datasets, SVM-based speaker verification performance is more dependent on the suitable selection of the T-norm dataset than that of the background dataset.

Comparing the results in Table 4 to those detailed in Tables 2 and 3, the refined disjoint datasets (Bkg=I_{250} and T-norm=T_{750}) were found to provide marginal performance improvements over the refined intersecting datasets (Bkg=B_{250} and T-norm=B_{1000}). This performance gain may also bring to light an underlying characteristic of background dataset refinement in that ranking of impostor examples may become more robust as the ratio of development SVMs in set S to the size of the full impostor set B is increase. Future work will investigate the impact that the size of B has on dataset refinement.

6 Conclusion

This study investigated the dependence of SVM-based classification performance to the selection of suitable background and T-norm datasets. The recently proposed background dataset refinement technique [5] was used to rank a large set of candidate impostor examples from which the top N highest-ranking observations were independently selected to form refined intersecting background and T-norm datasets. Evaluations were performed on both the NIST 2006 SRE development corpus and the unseen NIST 2008 SRE using a range of refined dataset sizes.

It was determined that SVM-based speaker verification classification performance is more sensitive to the selection of a suitable T-norm dataset than of background dataset selection. The best refined T-norm dataset, as determined by NIST 2006 development evaluations, provided substantial gains in both NIST 2006 and 2008 SRE irrespective of background choice. In contrast, the best refined background dataset offered only marginal performance improvements unless used in conjunction with the refined T-norm dataset, in which case maximum performance was obtained. Likewise, the refinement of disjoint background and T-norm datasets further demonstrated the high dependence of SVM-based speaker verification performance on the choice of T-norm dataset.

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

References

1. Kajarekar, S.S., Stolcke, A.: NAP and WCCN: Comparison of approaches using MLLR-SVM speaker verification system. In: Proc. IEEE ICASSP, pp. 249–252 (2007)
2. Stolcke, A., Kajarekar, S.S., Ferrer, L., Shriberg, E.: Speaker recognition with session variability normalization based on MLLR adaptation transforms. IEEE Trans. on Audio, Speech, and Language Processing 15, 1987–1998 (2007)
3. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digital Signal Processing 10(1), 42–54 (2000)
4. Burges, C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2), 121–167 (1998)
5. McLaren, M., Baker, B., Vogt, R., Sridharan, S.: Improved SVM speaker verification through data-driven background dataset selection. To be presented in Proc. IEEE ICASSP (2009)
6. Nation Institute of Standards and Technology: NIST speech group website (2006), `http://www.nist.gov/speech`
7. Campbell, W., Reynolds, D., Campbell, J.: Fusing discriminative and generative methods for speaker recognition: Experiments on switchboard and NFI/TNO field data. In: Proc. Odyssey, pp. 41–44 (2004)
8. Campbell, W., Sturim, D., Reynolds, D., Solomonoff, A.: SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In: Proc. IEEE ICASSP, pp. 97–100 (2006)
9. McLaren, M., Vogt, R., Baker, B., Sridharan, S.: A comparison of session variability compensation techniques for SVM-based speaker recognition. In: Proc. Interspeech, pp. 790–793 (2007)
10. Bonastre, J., Wils, F., Meignier, S.: ALIZE, a free toolkit for speaker recognition. In: Proc. IEEE ICASSP, pp. 737–740 (2005)
11. Chang, C., Lin, C.: LIBSVM: A library for support vector machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
12. NIST: The NIST Year 2008 Speaker Recognition Evaluation Plan (2008), `http://www.nist.gov/speech/tests/sre/2008/sre08_evalplan_release4.pdf`
13. Bengio, S., Mariéthoz, J.: A statistical significance test for person authentication. In: Proc. Odyssey, pp. 237–244 (2004)