IMPROVED SVM SPEAKER VERIFICATION THROUGH DATA-DRIVEN BACKGROUND DATASET SELECTION

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

The problem of background dataset selection in SVM-based speaker verification is addressed through the proposal of a new data-driven selection technique. Based on support vector selection, the proposed approach introduces a method to individually assess the suitability of each candidate impostor example for use in the background dataset. The technique can then produce a refined background dataset by selecting only the most informative impostor examples. Improvements of 13% in min. DCF and 10% in EER were found on the SRE 2006 development corpus when using the proposed method over the best heuristically chosen set. The technique was also shown to generalise to the unseen NIST 2008 SRE corpus.

***Index Terms*—** speaker recognition, data selection, support vector machines

1. INTRODUCTION

Most studies regarding SVM-based speaker verification have focussed on improving classification through the development of novel kernels and the optimisation of their associated parameters [1, 2, 3], however, a factor that can have a significant impact on classification performance is the choice of impostor set used in SVM training. In an SVM speaker verification system, a speaker model is commonly trained using a single utterance and a collection of negative or impostor observations, known as the background dataset. As the number of background examples significantly outweighs that of speaker examples, the SVM system relies heavily on the background observations in order to provide most of the observable discriminatory information. The background dataset must, therefore, consist of suitable impostor examples to ensure good classification performance.

The importance of selecting an appropriate background dataset to match the evaluation conditions has been highlighted in several recent studies [4, 5, 6]. The selection of a background dataset is often based on the knowledge of the broad characteristics expected in any impostor trials such as gender, language and the method of audio acquisition. Available data sources that satisfy these criteria can be utilised to compile various background dataset combinations which are then subject to development evaluations in order to heuristically select the most suitable impostor dataset. Although good performance can be obtained using this approach, it is not a systematic process, basing impostor selection on the performance of an entire set rather than analysing how much potential each impostor example offers to the background dataset.

This paper proposes an automated, data-driven technique for the selection of a suitable subset of impostor examples from a large and diverse set of examples for use as an SVM background dataset. A metric for the impostor suitability of an example is defined based on support vector selection frequency.

An example of heuristic background dataset selection is given in Section 2 along with a discussion on the drawbacks of the approach. Described in Section 3 is how support vector training can be exploited in the task of data selection followed by the proposal of the data-driven background dataset selection technique. Section 4 details the experimental configuration with results and discussions presented in Section 5.

2. HEURISTIC BACKGROUND DATASET SELECTION

Background dataset selection has traditionally been based on heuristics. This approach typically involves development evaluations in which various combinations of impostor datasets are used during SVM training. The background dataset contributing to the best performance is then regarded as the most suitable set of negative examples for the task of representing the impostor population.

As a starting point for further investigation, development evaluations were conducted on the English-only condition of NIST 2006 SRE using a GMM mean supervector SVM classifier employing nuisance attribute projection (NAP) [2]. T-norm score normalisation [7] was applied to all scores using the background dataset as the T-norm cohort. A detailed description of this configuration is described in Section 4.

2.1. Impostor Data Sources

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.

2.2. Development Evaluations

The results from development evaluations using a number of different background dataset configurations are detailed in Table 2. It can be seen from these results that the best performance was achieved when using the NIST 2004 corpus alone. For the remainder of the

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

[1] Selected randomly due to memory limitations restricting the full background dataset size to around 8000.

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

Table 2. T-normed results from English NIST 2006 SRE using different background datasets.

Background	**Min. DCF**	**EER**
NIST04	**.0135**	**2.82%**
NIST04 (English-only)	.0140	2.93%
NIST05	.0174	3.48%
Fisher	.0155	3.03%
SWB2	.0178	3.48%
NIST04 + NIST05	.0143	3.09%
NIST04 + Fisher	.0136	**2.82%**
NIST04 + SWB2	.0141	2.98%
Full Dataset	.0152	3.21%

paper, this set was defined as the heuristically-selected background dataset.

Combining the NIST 2004 dataset with impostor data from an alternate source resulted in the degradation of performance despite the significant increase in the number of impostor examples. Surprisingly, using only the English observations from the NIST 2004 dataset performed worse than the all-language NIST 2004 dataset despite development evaluations being performed on English-only trials.

The major shortcoming of the heuristic-based approach to background dataset selection is that the compilation of the candidate datasets lacks methodical structure and is based on very broad characteristics such as data source and language rather than on a per-example basis. Consequently, the apparent performance offered through heuristic-based selection relies on one of the candidate datasets being closely matched to the development evaluation conditions. Unless the suitability of each example for use in the background dataset is analysed, less appropriate impostor observations may unknowingly pollute the impostor population and ultimately prevent optimal classification performance from being achieved.

Although the results in Table 2 show that of the candidate datasets, the NIST 2004 dataset was the most suitable background dataset in the development evaluations, the classification performance when using alternate, single-sourced impostor datasets indicates that some degree of useful discriminative information was available in these sets. It seems reasonable, then, to assume that a subset of *useful* impostor examples may exist in each of these datasets. Without becoming a tedious and very time consuming task, the heuristic-based approach is unable to fully exploit these informative subsets of impostor examples.

3. DATA-DRIVEN BACKGROUND DATASET SELECTION

One approach to systematically selecting a background dataset for SVM training is to use a development dataset to drive the selection of impostor examples. Developing a method to rank all available impostor examples by their relevance to the background dataset will allow background dataset selection to be performed on a per-observation basis, thereby overcoming the shortcoming of the heuristic-based approach. A suitable criterion to perform this impostor observation ranking involves exploiting the information possessed by the support vectors of a trained SVM.

3.1. 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 [8]. 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* and are used to construct the separating hyperplane. Examples that are selected as support vectors hold a common property of being the most difficult to classify as they lie on or within the margin between classes. In contrast, those training examples that are not selected as support vectors provide no information in the training of the SVM.

The process of determining a subset of support vectors from the training observations can be considered a data selection process in which the most informative examples are chosen. 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. The resolution of the support vector frequency metric is dependent on the size of development dataset used for this purpose.

3.2. 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 impostor 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 background 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 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 background dataset is intended to be used in SVM training.

4. EXPERIMENTAL CONFIGURATION

The following experiments were developed with two objectives in mind. Firstly, to determine whether support vector frequency is reliable as an impostor suitability metric, and secondly, to determine the ability of the refined background dataset to generalise to the unseen data.

4.1. GMM-SVM System

SVM classification in the following experiments was based on GMM mean supervectors using the associated GMM mean supervector kernel [2]. 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 ALIZE/SpkDet package [10] distributed by LIA and is based on the the libSVM library [11]. Nuisance attribute projection (NAP) [2] was employed to reduce session variation with the 50 dimensions of greatest session variation being removed from all supervectors.

4.2. Evaluation Datasets

Large gender-dependent background datasets B were compiled from all available resources as listed in Table 1 and described in Section 2.1.

The gender-dependent development client dataset S used to calculate support vector frequencies was compiled from the English training and testing utterances in the 1conv4w condition of the NIST 2006 SRE, being the same development dataset used for heuristic-based selection in Section 2. Consisting of 1331 male and 1852 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 dataset generalised to unseen data. All NIST 2008 results were derived from condition 7 as specified in the official evaluation protocol which restricts trials to English spoken telephony data, matching the conditions found in the development dataset S.

T-norm score normalisation [7] was applied to all scores with results pooled after being evaluated on a gender-dependent basis. As an ideal set of T-norm models consists of a diverse range of impostor models, the background dataset was used as the T-norm cohort in this study. A leave-one-out approach was used to train these models. Consequently, the T-norm cohort was refined along with the background dataset.

5. RESULTS

5.1. Analysis of Support Vector Frequency

As an initial indicator as to whether the support vector frequency is an adequate metric to represent the suitability of an impostor observation, the 1000 highest ranked and 1000 lowest ranked supervectors were selected on a gender-dependent basis. Table 3 details the performance obtained using these background datasets compared to the full set of impostor examples.

The performance difference between evaluations using the 1000 examples of highest and lowest support vector frequency, detailed in Table 3, demonstrates that support vector frequency is an appropriate measure of the impostor example suitability. Despite consisting of the same number of observations, it is clear that the suitability of the background dataset to the evaluation corpus plays an important role in the performance of an SVM-based speaker verification system.

Through the removal of around 85% of the full set of impostor examples, the background dataset made up of the 1000 highest ranking observations provided a relative gain of 18% in DCF and 17% in EER, while using the 1000 lowest ranking observations demonstrated a loss of 18% in DCF and 8% in EER over the full background dataset. These statistics draw attention to two attributes associated with the choice of background dataset. Firstly, significant variation in classification performance is possible from background datasets of the same size and secondly, improved performance can be achieved by selecting backgound examples that are highly relevant to the evaluation conditions, even with a reduced overall quantity.

Table 3. Performance of T-normed scores from 1-sided, English NIST 2006 trials when using SVM background datasets refined by impostor support vector frequency.

Background Dataset	Min. DCF	EER
Full Dataset	.0152	3.21%
1000 Highest Frequency	.0125	2.65%
1000 Lowest Frequency	.0179	3.46%

5.2. Generalisation of Background Refinement Technique

Presented in Figure 1 is a plot of the min. DCF of both the NIST 2006 and 2008 evaluations as the background dataset was refined by removing the observations of the lowest support vector frequency determined using NIST 2006 data. Results are presented for both un-normalised and T-normed scores.

The plot demonstrates that the proposed background dataset refinement procedure generalises well to unseen data, however the conditions of the development data become more dominant in the refined impostor set when using less than around 2000 examples. T-normalisation appears to provide stable performance gains over un-normalised scores for the range of background dataset sizes despite the implicit refinement of the T-norm cohort (See Section 4.2).

Based on the plot in Figure 1, the best performance on the NIST 2006 corpus is found when using only 10-15% of the full set of impostor examples B in the background dataset. This presents an important finding that a large proportion of observations in the full dataset are not beneficial to the SVM background dataset and, in fact, contribute to a loss in performance even though they have originated from sources holding similar characteristics. The performance on the NIST 2008 corpus, on the other hand, is maximised when using between 15-50% of the full dataset indicating that if the full set is refined too extensively, the resulting impostor set may be too closely matched to the development corpus impeding its ability to generalise.

A sharp rise in the min. DCF curve in Figure 1 is observed as the size of the background dataset is reduced to less than around 500 observations. Although these small impostor sets consist of examples of the highest support vector frequency, the drop in performance draws attention to the necessity of adequate coverage of impostor space to prevent client SVMs from being under-trained [3].

Performance statistics of T-normed scores from evaluations using several different sizes of refined background datasets and the heuristically chosen background dataset of NIST 2004 data (as determined in Section 2) are detailed in Table 4. It can be observed that the heuristically selected background dataset provided a marginal 2% min. DCF improvement over the full dataset in the NIST 2008 SRE despite demonstrating a gains of more than 11% in both min. DCF and EER on the development corpus. The best performing evaluation on the NIST 2006 corpus was obtained when using the refined background dataset R_{500} providing a relative gain of 23% in min. DCF and 21% in EER over the full set of impostors. Using R_{500} as the background dataset in the NIST 2008 SRE provided more modest improvements of 6% in min. DCF and 3% gain in EER. These results demonstrate that the proposed background dataset refinement technique can sucessfully determine a more suitable set of impostor examples from a number of different sources than a set chosen through simple heuristic evaluations. Additionally, this superior performance was be achieved using less than 30% of the number

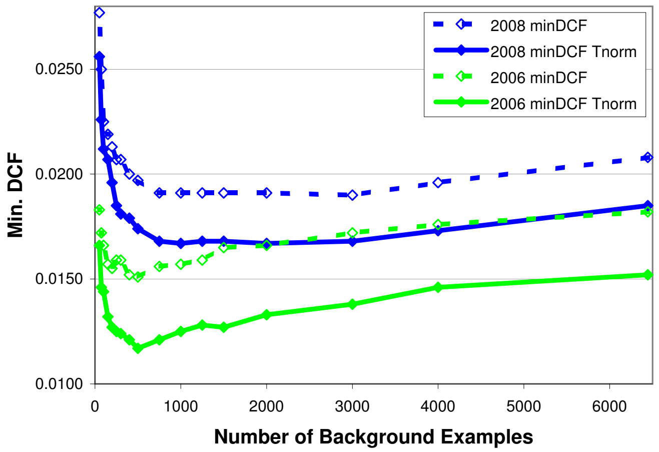

Fig. 1. Un-normalised and T-normed min. DCF on 1-sided, English NIST 2006 and 2008 SRE as the background dataset is refined.

Table 4. Min. DCF and EER obtained from T-normed scores on the 1-sided, English NIST 2006 and 2008 SRE when using full, heuristically determined and refined background datasets for SVM training.

Set	**2006**		**2008**	
	Min. DCF	**EER**	**Min. DCF**	**EER**
Full	.0152	3.21%	.0185	4.29%
Heuristic	.0135	2.82%	.0182	4.31%
R_{2000}	.0133	2.65%	**.0167**	**4.01%**
R_{1000}	.0125	2.65%	**.0167**	4.15%
R_{500}	**.0117**	**2.55%**	.0174	4.16%
R_{250}	.0125	2.60%	.0185	4.22%

of examples found in the heuristic set, increasing the computational efficiency of SVM training [3].

5.3. Database Contribution to Refined Background Dataset

The proportions of source databases that make up the best refined background dataset of 500 observations are detailed in Table 5 in terms of percentages. When compared to the full dataset, the NIST 2005 database maintains a steady contribution in the refined dataset while the contributions of both NIST 2004 and Fisher data increase to around 40%. A significant proportion of examples from the Switchboard 2 database appears to be unsuitable for use as background examples in the NIST 2006 SRE with most observations being removed through background dataset refinement.

In Section 2, it was shown that adding another dataset to the NIST 2004 dataset resulted in degraded performance. The figures in Table 5 indicate that the background dataset refinement technique was able to exploit a small subset of informative impostor examples from each of these alternative data sources, demonstrating the benefits of selecting a background dataset using a systematic, observation-based procedure rather than on a data source basis used in the heuristic-based approach.

6. CONCLUSION

Proposed was a novel, data-driven approach to the selection of a set of suitable impostor examples for use as the background dataset in SVM-based speaker verification systems through the refinement of a large and diverse dataset. Support vector selection frequency in the SVM training process was used to provide a measure of impostor suitability for each observation.

Table 5. Contribution of databases to full impostor set and refined background dataset of 500 highest ranking observations.

	Male		**Female**	
Data source	Full	R_{500}	Full	R_{500}
Fisher	31%	42%	26%	35%
Switchboard 2	31%	10%	26%	7%
NIST 2004	30%	41%	34%	42%
NIST 2005	8%	6%	14%	16%

The NIST 2006 SRE database was used to calculate support vector frequencies on a large set of impostor examples. Using a refined background dataset of the 500 highest ranking observations in the evaluation of the NIST 2006 corpus gave relative improvements of 23% in min. DCF and 21% in EER over the full background dataset demonstrating that the support vector frequency was an appropriate measure for the suitability of impostor example.

The background refinement technique was shown to generalise well to the NIST 2008 SRE where the refined background dataset was found to provide gains in both DCF and EER over a heuristically selected set of impostor examples.

7. REFERENCES

[1] L. Ferrer, K. Sonmez, and E. Shriberg, "A smoothing kernel for spatially related features and its application to speaker verification," in *Proc. Interspeech*, 2007, pp. 738–741.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, 2000.

[8] 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.

[9] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "A comparison of session variability compensation techniques for SVM-based speaker recognition," in *Proc. Interspeech*, 2007, pp. 790–793.

[10] J.F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *Proc. IEEE ICASSP*, 2005, pp. 737–740.

[11] C. Chang and C. Lin, *LIBSVM: A library for support vector machines*, 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.