Improved GMM-based Speaker Verification Using SVM-Driven Impostor Dataset Selection

Mitchell McLaren, Robbie Vogt, Brendan Baker, Sridha Sridharan

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

Abstract

The problem of impostor dataset selection for GMM-based speaker verification is addressed through the recently proposed data-driven background dataset refinement technique. The SVM-based refinement technique selects from a candidate impostor dataset those examples that are most frequently selected as support vectors when training a set of SVMs on a development corpus. This study demonstrates the versatility of dataset refinement in the task of selecting suitable impostor datasets for use in GMM-based speaker verification. The use of refined Z- and T-norm datasets provided performance gains of 15% in EER in the NIST 2006 SRE over the use of heuristically selected datasets. The refined datasets were shown to generalise well to the unseen data of the NIST 2008 SRE.

1. Introduction

Speaker verification systems depend on the selection of suitable impostor datasets to ensure good classification performance [1]. Impostor datasets are most commonly used for the purpose of score normalisation [2] in both Gaussian mixture model (GMM) and support vector machine (SVM) based configurations and the SVM background dataset. The importance of selecting appropriate impostor datasets to match evaluation conditions has been highlighted in several recent studies [3, 4, 5].

The selection of impostor datasets is often based on the knowledge of the broad characteristics expected in unseen impostor trials such as gender, language and the method of audio acquisition. Available data sources satisfying these criteria can then be used to form a combination of impostor datasets for use in development evaluations from which the most suitable impostor set can be selected. Although good performance can be obtained using this heuristic approach, the ranking of candidate datasets is based on the performance of the entire set rather than systematically assessing the potential that each candidate impostor example offers to the task at hand.

Data-driven background dataset refinement [1] was recently proposed as a technique to select the most suitable subset of examples from a large set of ranked candidate impostor observations for use as a *refined* SVM background dataset. Support vector selection frequency [1] was defined as the impostor suitability metric used to rank the candidate set. The refined background dataset was shown to provide significant improvements over the use of both the full candidate dataset and a background selected using the common heuristic-based approach to dataset selection. Further work discovered that refinement of a candidate T-norm dataset also provided significant performance improvements to SVM-based classification when using T-norm score normalisation [2, 3].

In this paper, the versatility of data-driven dataset refinement is demonstrated through its application to the selection of suitable impostor datasets for GMM-based speaker verification. A single candidate impostor dataset is ranked from which refined Z- and T-norm datasets are selected and evaluated on both development and unseen copora.

Section 2 briefly outlines the objectives of common score normalisation techniques and the importance of selecting suitable impostor datasets. Section 3 describes the data-driven dataset refinement technique for impostor dataset selection. Section 4 details the experimental protocol with results presented in Section 5.

2. Score Normalisation Techniques

Score normalisation techniques are an integral part of speaker verification systems as observed in the recent NIST speaker recognition evaluations (SRE) [6]. Score normalisation aims to counteract any statistical variations in classification scores by scaling all scores to a global distribution where a client- and test-independent classification threshold can be applied. Auckenthaler et al. [2] presented a study on the two most common forms of score normalisation: Zero and Test normalisation (Z-norm and T-norm respectively). Both techniques aim to produce a standard normal distribution for a set of non-target trial scores using

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

where the score s, obtained when comparing a given test segment to the speaker model, is normalised using mean μ_I and standard deviation σ_I obtained from an impostor score distribution. This impostor score distribution is estimated using an impostor dataset to achieve the following objectives:

- Z-norm aims to compensate for characteristics and biases of an individual speaker model by trialling the set of impostor examples against the client model.
- T-norm aims to compensate for characteristics and biases of a given test segment. This is performed by firstly training a speaker model from each impostor example in the impostor dataset. The test segment is then scored against each of these models to produce an impostor score distribution.

Commonly employed in GMM-based systems is the ZT-norm configuration in which Z-norm is applied to scores prior to T-norm. In this way, the benefits of both client-dependent and

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

test-dependent normalisation techniques are exploited to potentially provide improved classification performance.

The estimation of normalisation parameters, μ_I and σ_I, is reliable only if the impostor dataset is representative of the non-target trials in the intended evaluation corpus. If the impostor datasets could be tailored toward these conditions, the Z- and T-norm score transformations would be more effective.

3. Data-Driven Impostor Dataset Selection

Recently proposed in previous work was the data-driven background dataset refinement technique [1]. This technique uses a development dataset to drive the selection of impostor examples such that the resulting impostor dataset is representative of the conditions observed in the development set. The ability for this approach to tailor an impostor dataset toward a given set of conditions lends itself to the selection of GMM impostor datasets such as those mentioned in Section 2. The following study aims to determine whether the merit of data-driven refinement extends beyond the SVM domain.

Data-driven dataset refinement is based around an impostor suitability metric with which a candidate set of impostor examples is ranked allowing for dataset selection to be conducted on a single utterance-level basis. This overcomes the shortcoming of the heuristic-based approach in which the knowledge of broad characteristics of available data sources is used in determining candidate datasets rather than individually assessing the suitability of each impostor example for inclusion in the dataset. The support vector frequency was proposed in [1] as an appropriate ranking criteria that exploits the properties of the SVM training algorithms and the support vectors they select.

3.1. Support Vector Frequency

The support vector machine is a discriminative classifier trained to separate positive and negative classes in a high-dimensional space. A kernel is used to project input feature vectors into this high-dimensional space where a separating hyperplane is positioned to maximise the margin between the classes [7]. Support vectors are selected from the set of training examples to define the position of the separating hyperplane. The 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, training examples that are not selected as support vectors do not effect the hyperplane position.

For a single SVM model, 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 clients training data.

Extending this assumption over a large set of trained SVMs, the background examples that are more frequently selected as support vectors are likely to be relatively more important than those that are rarely selected. The *support vector frequency* of a candidate background example can, therefore, be defined as the total number of instances in which it is selected as a support vector when training a set of development client models. By this definition, the support vector frequency of an example provides a measure of its relative importance in the background dataset.

3.2. Background Dataset Refinement

Given a diverse set of speaker utterances or vectors B, compiled from a number of available resources, this candidate dataset can be refined into a suitable impostor dataset using a set of *development* client vectors S. The speakers 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 dataset.

As in the case of the heuristically selected dataset, the characteristics found in development dataset S are likely to influence the selection of the refined dataset R_N.

4. Experimental Configuration

The following experiments were developed with two objectives in mind; (1) to determine whether the SVM-based dataset refinement process is capable of selecting appropriate impostor datasets for use in GMM-based configurations, and (2) whether the refined dataset selected based on development evaluations generalises well to unseen data. All evaluations were performed using a GMM-UBM system with a GMM-SVM system used to train SVMs for the purpose of ranking impostor examples in the dataset refinement process.

4.1. GMM-UBM System

The GMM-UBM configuration used in this study utilises fully coupled maximum a-posteriori (MAP) adaptation and feature-warped MFCC features with appended delta coefficients, as described in [8]. An adaptation relevance factor of $\tau = 32$ and 512-component models were used throughout. Gender-dependent UBMs were trained using a diverse selection of 1818 utterances from both NIST 2004 and Switchboard 2 corpora.

The joint factor analysis model based on the approach of Kenny et al. [9] with elements described in [10] was employed to reduce the effects of session variation in the GMM-UBM system. The dimensionality of the gender-dependent speaker and session subspaces were set to 200 and 50 dimensions, respectively. This subspace was estimated using a diverse selection of the Switchboard II, NIST 2004 and 2005 SRE corpora.

4.2. GMM-SVM Refinement System

Training of SVM speaker models for the purpose of dataset refinement was based on GMM mean supervectors using the associated GMM mean supervector kernel [11]. The SVM implementation uses the open source ALIZE/SpkDet package [12] distributed by LIA and is based on the the libSVM library [13]. The GMM supervectors in this study were produced using the GMM system described in Section 4.1, however, a relevance factor of $\tau = 8$ was used and session compensation was not employed during GMM training. Instead, nuisance attribute projection (NAP) [11] was used in the SVM domain due to having a simple implementation while achieving comparable results to the factor analysis model [14]. As with session compensation in the GMM-UBM configuration, NAP was used to remove the

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

50 dimensions of greatest session variation from all supervectors during SVM training. This was a particularly interesting aspect of this study as the ranking of impostor examples was not conducted using supervectors from the same GMMs used in the GMM-UBM evaluations. The NAP transform matrix was trained using the same data as used to estimate the factor analysis subspace in the GMM domain.

4.3. Evaluation Datasets

As a starting point for refinement, large gender-dependent impostor 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 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.

The gender-dependent development client dataset, used to calculate support vector frequencies, was compiled from the English training and testing utterances in the 1conv4w condition of the NIST 2006 SRE. Consisting of 1331 vectors from 219 male speakers and 1852 vectors from 296 female speakers, this provided a moderate degree of resolution in the support vector frequency statistic. This dataset, along with the relevant test segments from the NIST 2006 SRE, were then used for development evaluations in order to determine the size of the refined datasets that corresponded to maximum performance.

The NIST 2008 SRE corpus was used to observe how well the refined impostor datasets generalised to unseen data. All NIST 2008 results were derived from condition 7 as specified in the official evaluation protocol [15] which restricts trials to English spoken telephony data, matching the conditions found in the development dataset.

In the following experiments, the minimum decision cost function (DCF) and equal error rate (EER) were used as performance criteria. The performance offered by the refined Z- and T-norm datasets was compared to that of heuristically selected datasets so as to provide a more thorough analysis of the dataset refinement technique. These heuristically selected datasets consisted of telephony speech from the NIST 2004 SRE corpus and were tuned and developed on performance statistics from the NIST 2005 and 2006 SRE corpora. A total of 376 utterances were used in both male and female Z-norm datasets and the T-norm datasets were made up of 125 male and 186 female utterances. The heuristically selected Z- and T-norm datasets contained observations from the same set of speakers, however, there were few utterances that existed in both impostor datasets.

5. Results

5.1. Development Evaluations

Support vector frequencies for each candidate impostor example were calculated using the NIST 2006 development dataset.

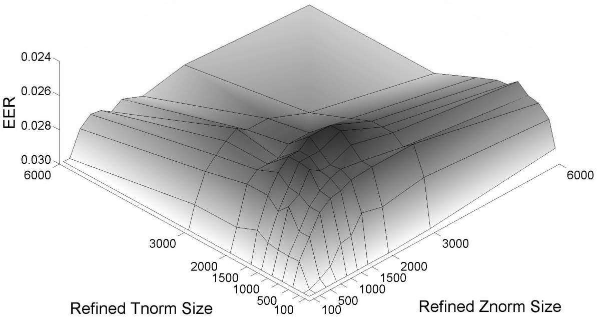

Figure 1: EER of ZT-normalised scores on NIST 2006 SRE when varying the size of the refined Z- and T-norm datasets.

Table 2: Min. DCF and EER obtained on 1-sided, English NIST 2006 SRE when using full, heuristically selected and refined impostor datasets for score normalisation.

	Male		Female		Combined	
	DCF	EER	DCF	EER	DCF	EER
Raw Scores						
	.0180	2.93%	.0180	3.83%	.0189	3.47%
Z-norm						
Full	.0187	2.95%	.0184	3.30%	.0187	3.25%
Heuristic	.0167	2.79%	.0173	3.20%	.0171	3.03%
Refined	.0165	2.52%	.0165	3.02%	.0167	2.92%
T-norm						
Full	.0167	2.79%	.0167	3.58%	.0171	3.20%
Heuristic	.0135	2.52%	.0167	3.66%	.0157	3.20%
Refined	.0111	1.97%	.0153	3.47%	.0137	2.93%
ZT-norm						
Full	.0122	2.52%	.0155	3.20%	.0145	2.87%
Heuristic	.0106	2.52%	.0141	3.13%	.0128	2.87%
Refined	.0093	1.86%	.0154	2.93%	.0132	2.44%

An exhaustive search was conducted to find the size of refined Z- and T-norm datasets that minimised the EER of the combined-gender ZT-normalised scores. The effect of varying the refined dataset sizes on the EER criterion can be observed in Figure 1. The darker regions of the plot designate improved performance. The EER criterion was minimised when using a Z-norm dataset of the 2000 highest-ranking impostor examples and a T-norm dataset of the top 1250 examples. As the refined Z- and T-norm datasets were both selected from the same ranked candidate impostor dataset, the T-norm dataset was, consequently, a subset of the larger Z-norm dataset.

Table 2 details the performance offered by the refined datasets when evaluated on a the NIST 2006 SRE. These results clearly demonstrate that data-driven background dataset refinement is suitable for the selection of score normalisation impostor datasets for use in the GMM domain. The refined datasets were found to provide superior performance to the heuristically chosen datasets in all Z- and T-norm performance statistics despite having their size selected to optimize *ZT-norm* EER performance. Dataset refinement appeared to provide greater benefits to T-norm scores than the Z-norm scores where relative improvements of 13% in min. DCF and 8% in EER were observed in the combined T-norm results over the heuristically selected datasets.

Consistent EER improvements were offered in the ZT-norm results of Table 2 through the use of refined datasets over the full or heuristically selected datasets. A relative improvement

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

Table 3: min. DCF and EER obtained on 1-sided, English NIST 2008 SRE when using heuristically selected and refined impostor datasets for score normalisation.

	Male		Female		Combined	
	DCF	**EER**	**DCF**	**EER**	**DCF**	**EER**
Raw Scores						
	.0230	4.14%	.0236	5.32%	.0240	5.13%
Z-norm						
Heuristic	.0250	4.55%	.0231	4.97%	.0243	4.80%
Refined	.0229	4.33%	.0212	4.82%	.0221	4.64%
T-norm						
Heuristic	.0207	4.05%	.0207	5.32%	.0209	4.64%
Refined	.0181	4.10%	.0178	5.04%	.0185	4.41%
ZT-norm						
Heuristic	.0198	5.19%	.0169	4.54%	.0181	4.64%
Refined	.0172	4.33%	.0172	3.93%	.0174	4.15%

of 15% in EER was observed in the combined ZT-norm results when using the refined datasets over the heuristic datasets while having a slightly worse min. DCF. This improvement coincides with the EER criterion minimised when selecting the refined dataset sizes.

An interesting observation from the results in Table 2 is that while the refined datasets gave superior min. DCF performance over the heuristic datasets in both Z- and T-norm results, the opposite was observed in the ZT-norm results (with the exception of the male results). This inconsistent min. DCF improvement is expected to be linked to the overlap of utterances in the Z- and T-norm datasets. While the same speakers were present in both the Z- and T-norm datasets in the heuristic case, there was minimal overlap between individual utterances. However, this was not the case for the refined datasets where the complete T-norm dataset was a strict subset of the Z-norm dataset. It is hypothesised that the reduced overlap of the heuristic datasets may have provided complementary information beneficial to ZT-normalisation that was not available in the refined datasets. Recent findings in dataset refinement for SVM-based classification support this hypothesis [3], however, further research is be required in order to determine whether the GMM-based score normalisation process benefits from impostor datasets that are not overlapping.

5.2. Generalisation of Refined Impostor Datasets

The refined Z- and T-norm datasets selected to minimise EER in the NIST 2006 development evaluations were used in the evaluation of the unseen NIST 2008 SRE. Table 3 details these results where it can be observed that the refined datasets generalised well to the unseen corpus. These datasets provided consistent performance improvements over the use of heuristically selected datasets with a gain of 4% in min. DCF and 11% in EER for the combined ZT-norm results. Similar trends were observed in the results of Table 3 to those of the NIST 2006 SRE in that the refined datasets provided greater gains in min. DCF over heuristic datasets when using Z- or T-norm independently, whereas the gains in the ZT-norm results were more apparent in the EER statistic.

6. Conclusion

The recently proposed data-driven background dataset refinement technique was applied to the selection of impostor datasets for the purpose of Z- and T-normalisation in a GMM-based speaker verification system.

The NIST 2006 SRE database was used to calculate the support vector frequencies of each example in a large set of candidate impostor examples. The size of the refined impostor datasets were selected to minimise the EER from ZT-norm results on the NIST 2006 SRE development corpus. These datasets provided a relative improvement of 15% in the ZT-norm EER on the development corpus over the use of heuristically selected datasets whereas the min. DCF showed a small drop in performance despite being improved for both Z- and T-norm individually.

The background refinement technique was shown to generalise well to the NIST 2008 SRE where the refined impostor datasets were found to provide gains in both DCF and EER over the heuristically selected datasets. These gains demonstrated that the SVM-based refinement technique was suitable for the selection of impostor datasets for use in GMM-based speaker verification.

7. References

[1] M. McLaren, B. Baker, R. Vogt, and S. Sridharan, "Improved SVM speaker verification through data-driven background dataset selection," *To be presented in Proc. IEEE ICASSP*, 2009.

[2] 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.

[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," *To be presented in Proc. ICB*, 2009.

[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] 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.

[6] Nation Institute of Standards and Technology, *NIST speech group website*, 2006, http://www.nist.gov/speech.

[7] 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.

[8] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey Workshop*, 2001, pp. 213–218.

[9] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor Analysis Simplified," in *Proc. IEEE ICASSP*, 2005, vol. 1, pp. 637–640.

[10] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

[11] 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.

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

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

[14] 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.

[15] NIST, *The NIST Year 2008 Speaker Recognition Evaluation Plan*, 2008, www.nist.gov/speech/tests/sre/2008/sre08_evalplan_release4.pdf.