QUT Speaker Identity Verification System for EVALITA 2009*

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. This document outlines the system submitted by the Speech and Audio Research Laboratory at the Queensland University of Technology (QUT) for the Speaker Identity Verification: Application task of EVALITA 2009. This submission consisted of a score-level fusion of three component systems, a joint-factor GMM system and two SVM systems using GLDS and GMM supervector kernels. Development and evaluation results are presented, demonstrating the effectiveness of this fused system approach.

Keywords: Speaker Identity Verification, EVALITA 2009.

1 System Overview

Three component systems were developed by QUT for this evaluation. These three systems are:

1. Joint Factor GMM-UBM system
2. GMM Supervector SVM system
3. GLDS SVM system

The submitted QUT system was the score-level fusion of these systems. Fusion was performed on the output scores using linear weights calculated through use of a logistic regression algorithm. This was performed using the FoCal toolkit [1]. Fusion weights were optimised for the minimum DCF on development evaluations.

The main focus of our development was on the training conditions *tc6* and *tc5* in which larger quantities of speaker training data from both channel conditions was available. Our systems, therefore, make little attempt to specifically deal with any channel-specific training conditions (*tc1* through *tc4*).

1.1 Development Protocol and Data

The EVALITA development protocols were derived from the set of 321 client access trials from 32 speakers for each of the testing conditions, 1 and 2 (denoted as *dev_ts1* and *dev_ts2*, respectively). In addition to these target trials,

* This research was supported by the Australian Research Council Grant DP0877835.

a speaker's impostor trials were made up of the non-target utterances in the development dataset. This resulted in a total of 5136 trials (2576 female, 2560 male) in each development evaluation condition.

The GMM-UBM configuration utilised data from the NIST SRE'04, NIST SRE'05 and Switchboard II corpora for system development as well as EVALITA *ubm* and *dev* data, as detailed below. In contrast to the GMM system, the development of the SVM subsystems was based solely on data sourced from EVALITA *ubm* speech.

2 Joint Factor GMM-UBM System

The acoustic subsystem was a GMM-UBM [2] system with a joint factor analysis model based on the approach of Kenny, et al. [3] with elements as described in [4] and [5]. The development of this system was geared toward consistent performance across all training/testing conditions.

2.1 Feature Extraction

Short-term cepstral feature vectors consisting of 12 MFCCs and 12 corresponding delta coefficients were used in this system. Before the features were extracted, the audio was bandpass filtered between 300Hz and 3.2KHz, followed by an energy based speech activity detection (SAD) process. Feature warping [6] was also applied using a 500-frame window. The application of feature warping in this system was not ideal as it was applied to each segment independently (rather than to all segments from a session). This lead to a situation with many segments having less active speech than the length of the 500-frame sliding window and, consequently, poor representation of the feature distribution.

2.2 Joint Factor Model

A joint-factor modelling approach very similar to [7] was adopted for this evaluation with low-dimensional subspaces for modelling both speaker characteristics and session/channel characteristics. The dimensianality of the gender-dependent speaker and session subspaces were set to 300 and 100 dimensions, respectively.

The speaker and session subspaces were estimated as follows. Gender dependent UBMs were trained based on all the NIST SRE'04 data with a selection of Switchboard II, Phase 2 and 3 data to increase the diversity of speakers represented. Based on the findings in [4] and [7], a "coupled" estimation method was used whereby the speaker subspace transform $\boldsymbol{V}$ was first fully optimised using an EM algorithm. A collection of telephony data from Switchboard II, and Mixer (SRE '04, SRE '05) were used for estimating $\boldsymbol{V}$. The session subspace transform $\boldsymbol{U}$ was then optimised again using an EM algorithm and using the previously trained speaker space. A selection of the EVALITA *ubm* data was used for this purpose, with concatenated excerpts from each session — of roughly the length of the *ts1* testing condition — combined to produce "sessions" from each of

the speakers. Around 5,000 of these "sessions" were produced per gender from the *ubm* data. Stacking of $\boldsymbol{U}$ matrices from EVALITA and NIST data was not found to be beneficial. Finally, $\boldsymbol{D}$ was estimated using data drawn from the *dev* speakers of roughly matched length to the *tc1*/*tc2* training conditions.

2.3 Scoring and Normalisation

Scoring was performed using a dot-product approximation of the log-likelihood ratio (LLR), as proposed in [8]. A dot-probuct was evalutaed between channel-compensated Baum-Welch statistics of the test utterance and the speaker model mean supervector, expressed as an offset from the UBM.

ZT-Norm was utilised for this system. T-Norm models were trained on EVALITA *ubm* data. Separate T-Norm lists were created for each of the training conditions, *tc1*–*tc6*, to mimic the quantity of active speech, number and source (PSTN or GSM) of sessions in each training condition. This resulted in 300 T-Norm models per gender in the *tc1*, *tc2* and *tc5* conditions and 90 per gender in the *tc3*, *tc4* and *tc6* conditions. Z-Norm segments also came from EVALITA *ubm* data and were formed in a similar manner, however, Z-norm utterance durations were matched to the testing length expected in the longer *ts2* test condition (using segments based on the *ts1* condition was found to be inferior due to the short length).

3 Support Vector Machine (SVM) System Commonalities

Two different SVM systems were used in this evaluation: GMM Supervector, and Generalised Linear Discriminant Sequence (GLDS) systems. Although the features between configurations differed, these systems had several common characteristics.

3.1 Intersession Variability Compensation

Nuisance attribute projection (NAP) [9] was applied to SVM systems to remove session variation in the SVM kernel space. The datasets used to train the gender-dependent projection matrices consisted of utterances from the EVALITA *ubm* data that contained more than 1.5 seconds of active speech. This resulted in approximately 7700 utterances from 30 speakers in each gender. Attempts were made to combine those training segments labelled as originating from the same session into fewer, longer utterances, however, this tended to reduce the benefits observed from the application of NAP. The 40 dimensions contributing the greater session variation were removed from all observations used in the GMM supervector SVM while 50 dimensions were removed from the GLDS kernel space.

3.2 SVM Background Dataset

Background examples were trained using the EVALITA *ubm* data. Each impostor example was trained using a minimum of S seconds of active speech from

the combination of speech segments from a single speaker. For the GMM supervector system, $S = 75$ for training conditions 5 and 6, while $S = 50$ for all other conditions. In the GLDS configuration $S = 100$ in all training conditions. Multiple examples were used per speaker where sufficient data was available. The channel and gender-dependent background datasets consisted of between 100 to 350 examples depending on the value of S and channel conditions.

The use of different lengths of training data was investigated using different values of S. Interestingly, performance was not maximised when matching the minimum active speech duration to the expected training or testing utterance duration. It is believed that a compromise was found between the quality and quantity of impostor examples in the background dataset with the S chosen.

3.3 Scoring and Normalisation

SVM-based classification scores were given by the distance that a test observations lies from a trained client hyperplane. In all configurations, the background dataset was used as the T-norm cohort. Consequently, the score normalisation cohorts matched the channel conditions observed in the training data. ZT-norm was employed only in the *tc5* and *tc6* conditions of the GMM-Svec configuration using appropximately 1000 Z-norm test segments formed from the EVALITA *ubm* data using $S = 25$.

4 GMM Supervector SVM System

The GMM supervector feature space was created from GMMs trained through MAP adaptation [2] from the UBM. The mixture component means were adapted using a relevance factor of $\tau = 8$ while the weights and variances remained constant. The MFCCs used in the adaptation process were previously described in Section 2.1 with the exception of the MFCCs used in training conditions 5 and 6. Here, feature warping was not employed as it was found to reduce performance.

The feature space of the SVM was based on the supervector formed from the concatenation of the adapted mixture component mean vectors. More specifically, the SVM feature space was established by taking the difference between the supervector of the concatenated Gaussian means of the UBM from the supervector formed from the means of the adapted GMM. In this evaluation, a single supervector was produced to represent a client using all their training segments.

The GMM supervector SVM configuration is based on the application of background-normalisation prior to the computation of the linear SVM kernel matrix [10]. In this technique, each dimension of the SVM feature space is normalised by the mean and standard deviation of the corresponding dimension of the observations in the background dataset. This normalisation process was performed subsequent to NAP.

Table 1. Min. DCF and EER obtained for each train-test combination on individual and submitted (Fused) systems.

Train		dev_ts1		dev_ts2		ts1		ts2	
Cond.	System	Min. DCF	EER	Min. DCF	EER	Min. DCF	EER	Min. DCF	EER
tc6	GMM-UBM	.1085	5.61%	.0529	2.25%	.1658	6.48%	.0893	2.75%
	Svec SVM	.0571	3.74%	.0046	0.62%	.2417	6.62%	.0549	1.78%
	GLDS SVM	.1347	6.85%	.0572	3.12%	.2876	11.98%	.1489	5.02%
	Fused	**.0367**	**2.18%**	**.0043**	**0.37%**	**.1168**	**4.59%**	**.0516**	**1.74%**
tc5	GMM-UBM	.2050	8.47%	.0979	4.67%	.2407	7.77%	.1395	4.25%
	Svec SVM	.1336	6.85%	.0503	2.49%	.2572	9.27%	.1013	3.48%
	GLDS SVM	.3121	11.21%	.1514	5.66%	.3551	14.87%	.2043	7.39%
	Fused	**.1220**	**4.67%**	**.0319**	**1.87%**	**.1610**	**6.03%**	**.0824**	**3.48%**
tc4	GMM-UBM	.2643	9.62%	.1171	5.60%	.3096	10.38%	.1826	6.72%
	Svec SVM	.2674	10.98%	.1412	4.92%	.3470	11.98%	.1761	5.99%
	GLDS SVM	.3735	14.07%	.2618	9.97%	.3845	16.75%	.2707	9.61%
	Fused	**.2063**	**8.69%**	**.1007**	**4.98%**	**.2700**	**9.75%**	**.1491**	**6.13%**
tc3	GMM-UBM	.2910	9.94%	.1312	5.99%	.2127	8.26%	.1234	4.11%
	Svec SVM	.3246	12.46%	.1768	7.73%	.3570	12.40%	.1545	5.99%
	GLDS SVM	.3723	15.58%	.2697	11.20%	.3901	16.23%	.2649	9.75%
	Fused	**.2467**	**9.03%**	**.1406**	**5.53%**	**.2270**	**8.15%**	**.0996**	**4.49%**
tc2	GMM-UBM	.3376	13.71%	.2165	9.97%	.3794	13.27%	.2529	9.27%
	Svec SVM	.3597	16.19%	.2632	8.73%	.4424	17.24%	.2452	9.61%
	GLDS SVM	.4973	19.31%	.3574	11.81%	.4536	20.27%	.3459	13.23%
	Fused	**.3594**	**12.77%**	**.2065**	**8.10%**	**.3575**	**11.78%**	**.2154**	**8.26%**
tc1	GMM-UBM	.3263	12.77%	.2057	8.40%	.2712	11.88%	.1900	6.76%
	Svec SVM	.3751	14.95%	.2467	9.03%	.4207	15.64%	.2500	8.64%
	GLDS SVM	.3888	19.31%	.3265	14.33%	.4770	20.24%	.3768	13.76%
	Fused	**.3191**	**11.84%**	**.1618**	**8.10%**	**.2919**	**11.15%**	**.1897**	**6.37%**

5 Generalised Linear Discriminant Sequence (GLDS) SVM System

The generalised linear discriminant sequence (GLDS) SVM configuration [11] was based on polynomial expansions. In this work, MFCC feature vectors of 24 dimensions (see Section 2.1) are utilised to produce the 4th degree polynomial basis terms resulting in an SVM feature space of 20475 dimensions.

Non-parametric rank normalisation [12] was employed in a linear SVM kernel. This technique operates by replacing each element of an input vector with its corresponding rank value when ranked against elements of the same index from a large set of vectors. The rank dataset for this task consisted of a subset of the NAP training utterances such that 40 utterances from each speaker were utilised. Rank normalisation was performed prior to NAP.

6 Results

Results of the individual and fused systems constituting this submission are presented in Table 1. It can be observed that significantly superior performance was offered by the GMM supervector SVM over the alternative classifiers in training conditions 5 and 6. In contrast, the GMM-UBM configuration tended to provide the best individual system performance in the single-channel training

conditions (PSTN for *tc4* and *tc2*; GSM for *tc3* and *tc1*). This was particularly true when limited testing data was available. While the individual system performance offered by the GLDS SVM was not always comparable to the other component systems, it was found to provide complementary information to the fusion process.

These results suggest that the GMM supervector SVM system found significant advantage when training client models using speech from both PSTN and GSM channels while the GMM configuration appeared more robust to channel-dependent training conditions with smaller amounts of training data.

Broadly speaking, the noted trends carried across well from the development conditions (*dev_ts1*, *dev_ts2*) to evaluation conditions (*ts1*, *ts2*). Furthermore, it can be observed that the performance of the GMM-UBM system suffered less from the transition to the unseen evaluation data. The fused evaluation results also show improvements in almost all conditions, often substantially so. It appears, however, that the thresholds chosen on the development data were not suited to the evaluation condition as the actual DCF values are quite poor. This result is most likely due to the sparsity of development results in the low miss operating region.

References

1. Brümmer, N.: FoCal: Tools for Fusion and Calibration of automatic speaker detection systems, available from: http://www.dsp.sun.ac.za/ nbrummer/focal/
2. Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted Gaussian Mixture Models. In: Digital Signal Processing, vol. 10, no. 1-3, pp. 19–41 (2000)
3. Kenny, P., Dumouchel, P.: Experiments in speaker verification using factor analysis likelihood ratios. In: Proc. of Odyssey04 - Speaker and Language Recognition Workshop, pp. 219–226 (2004)
4. Vogt, R., Baker, B., Sridharan, S.: Factor analysis subspace estimation for speaker verification with short utterances. In: Proc. of Interspeech 2008, pp. 853–856 (2008)
5. Vogt, R., Sridharan, S.: Explicit modelling of session variability for speaker verification. In: Computer Speech & Language, vol. 22, no. 1, pp. 17–38 (2008)
6. Pelecanos, J., Sridharan, S.: Feature warping for robust speaker verification. In: Proc. of A Speaker Odyssey, The Speaker Recognition Workshop, pp. 213–218 (2001)
7. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A study of interspeaker variability in speaker verification. In: Proc. of IEEE Transactions in Audio, Speech and Language Processing, vol. 16, no. 5, pp. 980–988 (2008)
8. Brümmer, N.: SUN SDV system description. In: NIST Speaker Recognition Evaluation Workshop Booklet (2008)
9. Solomonoff, A., Campbell, W., Boardman, I.: Advances in channel compensation for SVM speaker recognition. In: Proc. of the International Conference on Acoustics Speech and Signal Processing, pp. 629–632 (2005)
10. McLaren, M., Vogt, R., Baker, B., Sridharan, S.: A comparison of session variability compensation techniques for SVM-based speaker recognition. In: Proc. of Interspeech 2007, pp. 790–793 (2007)

11. Campbell, W.: Generalized linear discriminant sequence kernels for speaker recognition. In: Proc. of International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 161–164 (2002)
12. Stolcke, A., Kajarekar, S., Ferrer, L.: Nonparametric feature normalization for SVM-based speaker verification. In: Proc. of International Conference on Acoustics, Speech, and Signal Processing, pp. 1577–1580 (2008)