QUT NIST 2010 SRE System Description

Robbie Vogt, Brendan Baker, Mitchell McLaren, and Sridha Sridharan
Speech and Audio Research Laboratory, Queensland University of Technology
GPO Box 2434, Brisbane, AUSTRALIA, 4001.
Contacts: {r.vogt, bj.baker}@qut.edu.au

I. Introduction

A. System overview

Four main component systems were developed by QUT for this evaluation. The four systems are:

1) Joint factor GMM-UBM system
2) Superfactor system
3) GMM Supervector SVM System (MFCC Features)
4) GMM Supervector SVM System (MFDP Features)

In line with previous evaluations, the main focus of our development was on the English-only telephone condition. Microphone data was included in session compensation models (NAP etc.) in order to account for the variation encountered in this form of data.

The QUT primary system used for the *core–core* task used the output from all four systems. Results for the primary system are designated QUT_1 in the submission. Alternate systems were submitted for the *core-10sec*, and *10sec-10sec* conditions that utilised scores produced by the Joint factor GMM-UBM subsystem only.

B. Evaluation tasks performed

Three main evaluation tasks were addressed with QUT systems. The *core-core*, *core-10sec*, and *10sec-10sec* conditions. In addition, a system was evaluated for the *extended core-core* condition.

C. Development data and procedures

The NIST 2008 SRE data and protocol were used as the development database and testing protocol. All system tuning and fusion training was performed using the *short2–short3*, *long–short3* and *long–long* conditions of this data set.

Data from the SRE '04, SRE '05 and Switchboard II corpora were utilised for background model training, score normalisation and other system development components as detailed below. Data from the Fisher corpora was also used for SVM background training examples.

II. Joint Factor GMM-UBM System

The acoustic subsystem was a GMM-UBM [1] system with a joint factor analysis model based on the approach of Kenny, et al. [2] with elements as described in [3] and [4], and dot-product scoring [5].

A. Feature Extraction

Short-term cepstral feature vectors consisting of 13 MFCCs, including c0, and the corresponding delta coefficients were used in this system. Before the features were extracted, the audio was band filtered between 300Hz and 3.2KHz, followed by an energy based speech activity detection (SAD) process. After extracting the cepstral features, feature warping [6] was also applied using a 500 frame window.

Unlike previous years, no effort was made to perform echo cancellation. Additionally, NIST's voice activity detector (*speech*) was not used. While this caused some mismatch in the signal processing of the development data from various sources (e.g. SRE'04 data, Switchboard, etc.), removing these steps greatly simplified feature extraction. It was also found that echo cancellation significantly degraded the room microphone recorded speech.

Apart from simplifying the signal processing, no specific allowance was made to account for the interview conditions of the evaluation. We expect that this will have negative consequences on system performance, however, implementing and testing a speaker diarisation front end, as required by this data, was far too costly an exercise to enter into, and falls outside the scope of "speaker recognition evaluation."

B. Joint Factor Model

A joint-factor modelling approach [2] was once again utilised 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 400 and 100 dimensions, respectively. Residual relevance MAP adaptation was also applied. The JFA hyperparameters $\boldsymbol{V}$, $\boldsymbol{U}$ and $\boldsymbol{d}$ were estimated following the procedure presented in [7].

An additional session subspace was estimated on auxiliary microphone data drawn from the SRE '05 corpus. A 100-dimensional session subspace transform was created as an orthogonalised combination of the first 50 dimensions of both the telephone and auxiliary mic session transforms.

1) Model Training: Gender dependent, 512-component UBMs were trained based on all of the SRE '04 data.

Individual speaker models were adapted from the UBMs using a MAP process that simultaneously optimised both the speaker and session factors as well as the relevance MAP adaptation. For efficiency/practicality, this simultaneous optimisation used an approach similar to the Gauss-Seidel method for solving linear systems as described in [3].

C. Scoring and Normalisation

Scoring was performed using Brümmer's dot-product approximation of the log-likelihood ratio between the speaker model and the UBM, incorporating channel compensation [5].

Gender-dependent ZT-Norm was utilised for this system. The same set of utterances, drawn from SRE'04 and SRE'05, were used for both Z-Norm and T-Norm. A core set of 420 female and 307 male telephone segments were used in all cases, with the additional inclusion of 133 female and 113 male microphone segments in microphone-recorded conditions. For example, an interview-train, telephone-test trial used *tel* + *mic* for T-Norm, but *tel*-only for Z-Norm.

D. Reverse Scoring

Reverse scoring was also performed for the *core-core* and extended *core-core* conditions, whereby the evaluation protocol is performed in the reverse sense. In the reversed scoring scenario, the test utterances and training utterances are switched such that models are created from the test utterances while these models are scored using the training utterances.

E. Modifications for the 10sec-10sec Condition

To account for the differences of very limited train and test data in the *10sec-10sec* condition, alterations were made to the session factor loading matrix and the score normalisation cohorts. For both the session factor training and normalisation, the utterances used for these tasks were truncated to approximately 10 seconds of active speech to match the evaluation conditions. All data was telephone-only. No other changes were made.

F. Modifications for the core-10sec Condition

For the mismatched-length condition of *core-10sec*, a combined session factor loading matrix was produced using a combination of the full-length and truncated session factor matrices. The 50 leading dimensions from each matrix were concatenated and orthogonalised to produce the final 100-dimensional session subspace. No other changes were made and the standard telephone-only normalisation cohort was used in this condition.

III. Superfactor GMM-UBM System

A superfactor variant [8] of the the above Joint Factor GMM-UBM system was also included in the primary system for the *core-core* condition. With this variant, a number of estimates of the speaker factors, $\boldsymbol{y}$, are produced for each model using the Baum-Welch statistics corresponding to a subset of the UBM mixture components; these $\boldsymbol{y}$ estimates are then concatenated to form a supervector of speaker factors. The Superfactor system uses a symmetric scoring approach by performing a dot-product between the speaker superfactors of the training and test utterances. The 512 components were divided into 32 splits for this evaluation. In all other respects, the Superfactor configuration mirrored the standard Joint Factor GMM-UBM system.

IV. Support Vector Machine (SVM) System Commonalities

Two different SVM systems were used in this years evaluation for the *core-core* condition: A GMM Supervector SVM system based on MFCC feature extraction, and an alternate GMM Supervector system utilising Mel-frequency Delta Phase (MFDP) features.

A. Intersession Variability Compensation

Nuisance attribute projection (NAP) was applied to both SVM systems. Gender-dependent NAP projection matrices were trained on both telephone and auxiliary microphone data sets respectively. A combined projection matrix was then created as an orthogonal combination of the first N dimensions of both the telephone and auxiliary mic transforms. The datasets used to train the gender-dependent projection matrices were based on a collection of utterances from male and female speakers from NIST 2004, NIST 2005 and Switchboard 2 data.

B. Model Generation

The process for training a speaker model consists of creating a target speaker supervector from a training utterance and a set of impostor supervectors from a group of out-of-set speaker utterances (also refered to as the background dataset). We used gender-dependent background datasets derived from NIST 2004 and NIST 2005 data consisting of 487 male and 467 female utterances. Both SVM systems were based on the linear kernel. An optimal separating hyperplane is then calculated. The speaker model is represented by a normal vector to the hyperplane and an offset.

C. Scoring and Normalisation

In testing, the test supervector is created in the same manner as the training supervector. A dot product between this vector and the normal to the training model hyperplane with the offset is used to provide the speaker score. Scores were further normalized using both Z- and T-Norm datasets [9] which were constructed from the NIST 2004 and 2005 datasets.

V. GMM Supervector SVM System (MFCC)

A. Feature Extraction

The GMM supervector feature space is created from Gaussian Mixture Models (GMMs) trained through maximum-a-posteriori (MAP) adaptation [10] from a Universal Background Model (UBM) to the features of a speaker's utterance. The mixture component means are adapted using a relevance factor or $\tau = 8$ while the weights and variances remain constant. The feature space of the SVM is based on the supervector formed from the concatenation of the adapted mixture component mean vectors. More specifically, the SVM feature space is established by taking the difference between the supervector of the concatenated Gaussian means of a UBM from the supervector formed from the means of the adapted GMM. The 80 greatest dimensions contributing to session variation were removed from all observations using NAP (40 dimensions trained on each of both telephony and microphone data, stacked and orthognalised)

B. SVM Kernel

The GMM supervector SVM configuration is based on the application of background-normalisation prior to the computation of the linear SVM kernel matrix [11]. 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.

VI. GMM Supervector SVM System (MFDP)

The MFDP GMM Supervector system differs only from the MFCC implementation in the type of features extracted.

A. Feature Extraction

This system utilises a newly derived Mel-frequency Delta Phase (MFDP) feature set [12]. These mel-frequency cepstral coefficients are derived from the Delta-Phase Spectrum. These features were proposed as a simple phase domain representation that lend to consistent comparison across multiple frames and sequences, while circumventing issues associated with phase wrapping. The feature extraction used follows that described in [12], with a 250ms frame length and 10ms step size. For the implementation used in this evaluation, the first 12 cepstral coefficients with appended deltas were extracted, resulting in a 24-dimensional MFDP feature vectors.

Feature extraction was once again followed by GMM training through MAP adaptation using a relevance factor $\tau = 8$. The feature space of the SVM is based on the supervector formed from the concatenation of the adapted mixture component mean vectors. For the MFDP system, the 60 greatest dimensions contributing to session variation were removed from all observations using NAP (30 dimensions trained on each of both telephony and microphone data, stacked and orthognalised)

B. SVM Kernel

The GMM supervector SVM configuration is based on the application of background-normalisation prior to the computation of the linear SVM kernel matrix [11]. 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.

VII. Fusion and Development

In order to obtain an overall score for each test segment, the scores from the various subsystems were fused. Fusion was performed on the output scores using linear weights calculated through use of bilinear logistic regression. This was performed using the Bilinear FoCal toolkit provided by Niko Brümmer.

The side information used in the *core-core* and extended conditions encoded combinations of training and testing *source*. We had three distinct categories of source for both training and testing, being *mic*, *tel* and *long* defined as follows:

- *mic* includes all data recorded through room microphone channels, including both conversational style and interview style data, excluding long interview recordings (form the `interview/8min` directory).
- *tel* includes all data recorded over telephone channels.
- *long* includes interview segments with long durations (that is, from the `interview/8min` directory).

The *core-core* condition exhibits 7 of the possible 9 unique combinations of these conditions (omitting *tel-long* and *tel-mic*). We used a development protocol comprising a combination of *short2-short3*, *long-short3* and *long-short3* in the reverse sense from the SRE 2008 data. This development set covered all combinations exposed in the SRE 2010 *core-core* condition.

Our primary submission, QUT_1 was a fusion of 5 components; JFA, Reverse JFA, Superfactors, GSV-SVM with MFCC features, GSV-SVM with MFDP features.

The QUT_2 alternate system was a fusion of the same components but excluding the MFDP feature based GSV-SVM system. This excluded system appeared to have a anomalous score distribution for the evaluation data and we suspect some processing errors may have occurred.

The QUT_3 alternate system comprises only the forward and reverse JFA systems. This system was included for direct comparison between the *core-core* and extend *core-core* conditions as it was not practical to run all 5 component systems of QUT_1 for the extended condition.

The submitted systems for the *core-10sec* and *10sec-10sec* conditions consist of only the forward JFA system, with logistic regression applied to improve score calibration. These systems were designated QUT_1.

References

[1] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 19–41, 2000.

[2] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in *Odyssey: The Speaker and Language Recognition Workshop*, 2004, pp. 219–226.

[3] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.
[4] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech*, submitted, 2008.
[5] A. Strasheim and N. Brümmer, "SUNSDV system description: NIST SRE 2008 evaluation," in *NIST SRE Workshop*, 2008.
[6] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *A Speaker Odyssey, The Speaker Recognition Workshop*, 2001, pp. 213–218.
[7] P. Kenny, N. Dehak, V. Gupta, P. Ouellet, and P. Dumouchel, "A new training regimen for factor analysis of speaker variability," in *ICASSP*, 2008.
[8] N. Scheffer and R. Vogt, "On the use of speaker superfactors for speaker recognition," in *ICASSP*, 2010.
[9] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 42–54, 2000.
[10] D. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Eurospeech*, vol. 2, 1997, pp. 963–966.
[11] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "A comparison of session variability compensation techniques for SVM-based speaker recognition," in *Interspeech 2007*, 2007, pp. 790–793.
[12] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum: A consistent representation of the short-time phase of speech," *submiitted to IEEE Trans. on Audio, Speech and Language Processing*, 2010.