A Comparison of Session Variability Compensation Techniques for SVM-Based Speaker Recognition

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 paper compares two of the leading techniques for session variability compensation in the context of GMM mean supervector SVM classifiers for speaker recognition: inter-session variability modelling and nuisance attribute projection. The former is incorporated in the GMM model training while the latter is employed as a modified SVM kernel. Results on both the NIST 2005 and 2006 corpora demonstrate the effectiveness of both techniques for reducing the effects of session variation. Further, system- and score-level fusion experiments show that the combination of the two methods provides improved performance.

Index Terms: Support Vector Machines, Inter-Session Variability Modelling, Nuisance Attribute Projection, GMM

1. Introduction

Gaussian mixture models (GMMs), particularly within the GMM-UBM configuration [1], have proven to be an effective approach to speaker recognition. In such a system, the GMM is a generative model that is trained to best represent the distribution from which observed data was produced.

While the GMM-UBM configuration has become a standard approach, the introduction of the support vector machine (SVM) has motivated research into the benefits of discriminative classification for speaker verification. Automatic speaker verification systems incorporating SVMs have resulted in performance comparable, and in some cases superior to the GMM-UBM method [2, 3].

A significant amount of focus has been given to the fusion of these generative and discriminative techniques. Campbell et al. demonstrated the potential in this approach by proposing a GMM mean supervector SVM classifier [3]. In this configuration, GMM mean supervectors — formed through the concatenation of adapted GMM component means — are the input features to an SVM classifier.

Both the GMM-UBM and SVM based approaches to speaker verification suffer significant performance degradation due to the effects of channel and session variation. These variations occur when the channel and environmental conditions during acquisition of training and testing utterances differ. Techniques have been developed to combat this issue for GMM-UBM systems and have more recently been tailored for SVM-based systems. Two of the prominent methods that have been proposed to compensate for session variation in GMM mean supervector SVM systems are inter-session variability modelling [4] and nuisance attribute projection [5].

The first approach, inter-session variability (ISV) modelling, attempts to model the effects of the session differences in the GMM modelling process as a mean offset constrained to a low-dimensional session subspace [4]. In this approach, the speaker model parameters and the session offset are simultaneously optimised according to maximum a-posteriori (MAP) criteria. ISV modelling can be incorporated into an SVM system by using the mean supervectors extracted from the resulting session-compensated GMMs.

Alternatively, nuisance attribute projection (NAP) can address the issue of session variation in the SVM kernel. Rather than modifying the GMM modelling process, NAP modifies the SVM kernel to project observations onto a subspace that is more resistant to session variation [6]. This subspace is determined by removing the dimensions that are dominated by unwanted or nuisance variation, such as session variation.

This paper compares the benefits of modelling session effects through the generative ISV modelling approach, to the discriminative modelling approach of NAP, in a GMM mean supervector SVM system. The combination of the two technologies into a single system is also investigated along with the fusion of scores from the two individual systems.

Details of the GMM mean supervector SVM speaker verification system are first presented in Section 2. Section 3 describes the ISV modelling and NAP approaches to modelling session variation and discusses their similarities and differences. Sections 4 and 5 detail the experimental configuration and results when evaluated using both the NIST 2005 and NIST 2006 corpora.

2. A GMM Mean Supervisor SVM System

The GMM mean supervisor SVM system combines the idea of representing acoustic observations in terms of adapted GMM mean vectors with discriminative SVM classification. The mean supervectors provide a convenient method of mapping an utterance from a variable-length utterance to a fixed-dimension vector as required for use within an SVM classifier.

2.1. Support Vector Machines

The motivation for support vector machines (SVM) was to perform classification by mapping observations to a high-dimensional, discriminative space while maintaining good generalisation characteristics [7]. SVM training involves the positioning of a hyperplane in the high-dimensional space such that the maximum margin exists between classes. The term support vectors refers to the training vectors which are located on or between the class boundaries and, as a result, contribute to the positioning of the separating hyperplane. A kernel function
Figure 1: Stages involved in a GMM mean supervector SVM speaker verification system.

\[ K(X_a, X_b) = \phi(X_a) \cdot \phi(X_b) \] is used to compare observations in the high-dimensional space to avoid explicitly evaluating the mapping function \( \phi(X) \).

### 2.2. GMM Mean Supervectors

In order to produce a GMM mean supervector, a GMM must first be trained. Commonly used is the GMM-UBM configuration in which MAP adaptation is employed to adapt only the means of the universal background model (UBM) to represent a set of acoustic observations.

The GMM likelihood function is given by

\[ g(x) = \sum_{c=1}^{C} \omega_c \mathcal{N}(x; \mu_c, \Sigma_c), \tag{1} \]

where \( \omega_c \) are the component mixture weights, \( \mu_c \) the means, and \( \Sigma_c \) the covariances of the Gaussians. A mean supervector can be obtained from an adapted GMM by concatenating each of the component mean vectors, \( \mu = [\mu_1^T \ldots \mu_C^T]^T \).

### 2.3. Implementation of the System

The flow of data through a GMM mean supervector SVM speaker verification system is shown in Figure 1.

During speaker training, each available utterance of the speaker is first used to train a GMM through MAP adaptation from a UBM. Supervectors are then extracted from this set of adapted GMMs and used as examples of the speaker in SVM training to produce a speaker model. For an utterance \( X \), the mapping function is therefore \( \phi(X) = \mu \), where \( \mu \) is the adapted GMM mean supervector. In this arrangement, the supervector can be viewed as the information link between the generative and discriminative modelling processes. As SVM training is discriminative, examples of the non-speaker class are also required; this role is fulfilled by a set of supervectors from a representative background population of speakers.

In the testing phase, the same procedure for producing supervectors in training is used to generate a supervector for each test utterance. Comparing a test supervector to a trained SVM speaker model produces a classification score that is the distance of the test vector from the SVM hyperplane.

### 3. Approaches to Session Variability Compensation

This section describes two of the leading approaches to modelling session variation in the GMM mean supervector SVM system: ISV modelling and NAP. Although each method is performed in a different domain — the GMM modelling domain and the SVM expansion space — the techniques share common characteristics.

#### 3.1. Inter-session Variability Modelling

Attempts to directly model session variability in GMM-UBM based speaker verification systems have provided significant performance improvements in telephony environments [4, 8]. The purpose of inter-session variability (ISV) modelling is to introduce a constrained offset to the speaker’s GMM mean vectors to represent the effects brought about by the session conditions. In other words, the Gaussian mixture model that best represents the acoustic observations of a particular recording is the combination of a session-independent speaker model with an additional session-dependent offset. This can be represented in terms of the GMM mean supervectors as,

\[ \mu_h(s) = m + y(s) + Uz_h(s). \tag{2} \]

Here, the speaker \( s \) is represented by the offset \( y(s) \) from the speaker-independent (or UBM) mean supervector \( m \). To represent the conditions of the particular recording (designated with the subscript \( h \)), an additional offset of \( Uz_h(s) \) is introduced where \( z_h(s) \) is a low-dimensional representation of the conditions in the recording and \( U \) is the low-rank transformation matrix from the constrained session variability subspace to the GMM mean supervector space.

A GMM speaker model is trained through the simultaneous optimisation of the model parameters \( y(s) \) and \( z_h(s) \), \( h = 1, \ldots, H \) over the speaker’s training utterances. All model parameters are optimised according to the maximum a posteriori (MAP) criterion. The speaker offset \( y(s) \) has a prior as described by Reynolds [1] while the prior for each of the session factors \( z_h(s) \) is assumed to belong to a standard normal distribution, \( \mathcal{N}(0, I) \). An efficient procedure for the optimisation of the model parameters is described in [4].

This modified GMM training procedure with ISV modelling can be used in place of standard MAP adaptation GMM training in the SVM system of Figure 1 for both training and testing. The supervectors presented to the SVMs are in this way compensated for session effects.

#### 3.2. Nuisance Attribute Projection

Nuisance attribute projection (NAP) has been successful in combating session effects in the SVM kernel [5]. NAP accomplishes this with a modified kernel matrix to perform projection as opposed to the GMM mean offsets used in ISV modelling. NAP essentially finds a new kernel expansion space on which observations are projected, providing greater resistance to session and channel effects. The modified kernel is formulated as,

\[ K(X_a, X_b) = P \phi(X_a) \cdot P^T \phi(X_b) \tag{3} \]

where the projection matrix \( P = I - U_n U_n^T \). The purpose of the projection matrix is to remove the nuisance directions encoded in \( U_n \), thereby minimising the average distance between input vectors from the same speaker — that is, removing variation between different sessions of the same speaker.

As described in [5, 6], the training of the projection matrix \( P \) involves finding the matrix \( U_n \). Given the matrix \( A \) whose columns contain all the supervectors from a projection training dataset, \( U_n \) can be found by determining the set of eigenvectors, \( V \) with the largest eigenvalues satisfying

\[ (\text{diag}(W^2) - W)Kv = \lambda v, \tag{4} \]
where 1 is a column vector of ones, \( K = AA^T \) is the kernel evaluation matrix and \( W \) is the weight matrix indicating observations in the training dataset of the same speaker. \( U_n \) is then given by \( U_n = AV \). Readers are directed to [5, 6] for a more thorough explanation of the NAP kernel and training procedures.

NAP SVM speaker models are trained using the same procedure as the standard GMM mean supervector SVM system while employing the modified kernel with projection matrix \( P \) from (3). With reference to the flowchart in Figure 1, the NAP kernel is employed during both the SVM training and SVM testing stages.

### 3.3. Relationship between ISV modelling and NAP

In many ways, ISV modelling and NAP are very similar. This similarity is highlighted by noting that both the final ISV modelling and NAP compensated vectors \( y \) can be expressed in the form

\[
y = \mu - Uz.
\]

(5)

It is obvious from this formulation that both techniques compensate for session variation by removing the estimated influence of the session conditions, \( z \), on the mean supervector. In both cases, \( z \) is assumed to be linear in nature and restrained to a low-dimensional subspace, as defined by the projection matrix \( U \). The differences between the approaches arise in how \( z \) and \( U \) are estimated.

For ISV modelling, a generative approach is taken where both \( y_{\mathrm{isv}} \) and \( z_{\mathrm{isv}} \) are simultaneously optimised to maximise the \textit{a posteriori} probability of the observed feature vectors. As described in (2), \( \mu_{\mathrm{isv}} \) is essentially a by-product defined as the sum of \( y_{\mathrm{isv}} \) and \( Uz_{\mathrm{isv}} \).

In the case of NAP, estimating the effect of the session conditions does not rely on a probabilistic approach and does not directly make use of the acoustic observations: The session conditions are simply calculated by projecting the mean supervector \( \mu_{\mathrm{nap}} \) onto the session subspace,

\[
z_{\mathrm{nap}} = U^T \mu_{\mathrm{nap}}.
\]

(6)

It is also noteworthy that the projection matrices are very similar, in fact the ISV modelling projection \( U \) is initialised to be equivalent to the NAP matrix \( U_n \), but is then refined through an iterative process to further optimise it’s probabilistic representation of the training dataset. There is some evidence that this optimisation isn’t necessary [8] in which case the same matrix can be used for both techniques.

### 4. Experimental Configuration

The session variation compensation techniques described in Section 3 were evaluated using the NIST 2005 and NIST 2006 speaker recognition evaluation (SRE) corpora of telephony speech. The experiments focussed on the 1-sided training and testing condition detailed in the NIST evaluation plans [9]. All results are for the \textit{common evaluation} condition (restricted to English utterances) with both genders pooled.

The GMM training used in this study utilises fully coupled MAP adaptation and feature-warped MFCC features with appended delta coefficients, as described in [10]. An adaptation relevance factor of \( \tau = 8 \) and 512-component models are used throughout. Gender-dependent UBMs were trained using a diverse selection of 1818 utterances from both NIST 2004 and Switchboard 2 corpora.

The background-scaling linear kernel [11] was implemented using LibSVM [12] for SVM training and classification. This kernel uses statistics based on the background dataset to scale each dimension of the input supervectors to have unit variance thus allowing each dimension to contribute equally in the training of the SVM. The kernel is formulated as,

\[
K(X_a, X_b) = (\mu_a - m)^T B^{-1} (\mu_b - m),
\]

(7)

where \( B \) is the diagonal covariance matrix of the background dataset and \( m \) the UBM mean supervector.

In order to provide negative examples for SVM training, two gender-dependent background datasets were collected, each containing a selection of 2000 English utterances from unique speakers drawn from the LDC Fisher Corpus parts 1 and 2 [13].

The subspace matrix \( U \) for ISV modelling and the NAP kernel projection matrix \( U_n \) were both trained from a projection dataset. This dataset was also extracted from the LDC Fisher Corpus parts 1 and 2 and consisted of a total of 1344 male speakers and 1500 female speakers distinct from the background dataset each with three or more English utterances.

For both session variation compensation approaches, the 50 dimensions demonstrating the greatest session variability were retained for these experiments. The NAP projection matrix \( U_n \) was trained according to the procedure described in [6] while the ISV modelling matrix \( U \) was trained according to [4] using 10 EM iterations.

### 5. Results

The session variation compensation techniques are compared to the baseline GMM mean supervector SVM system in Figure 2 for the 1-sided train and test (1conv4w-1conv4w) common evaluation condition of the NIST 2005 SRE corpora. Table 1 details the equal error rate (EER) and minimum decision cost function (DCF) of each system when evaluated using both NIST 2005 and 2006 corpora.
Table 1: Minimum DCF and EER results for 1-sided trials of GMM mean supervector SVM systems evaluated using the NIST 2005 and 2006 SRE corpora.

<table>
<thead>
<tr>
<th>System</th>
<th>NIST ’05</th>
<th>NIST ’06</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EER</td>
<td>Min. DCF</td>
</tr>
<tr>
<td>Baseline</td>
<td>6.43%</td>
<td>.0248</td>
</tr>
<tr>
<td>NAP</td>
<td>5.40%</td>
<td>.0188</td>
</tr>
<tr>
<td>ISV</td>
<td>4.97%</td>
<td>.0184</td>
</tr>
<tr>
<td>NAP ISV</td>
<td>4.97%</td>
<td>.0180</td>
</tr>
<tr>
<td>Fused</td>
<td>4.93%</td>
<td>.0170</td>
</tr>
</tbody>
</table>

Results show that the ISV modelling method provides consistent results with a reduction in EER of 29% and 22% over baseline results for the NIST 2005 and NIST 2006 corpora respectively. The performance of NAP is similar to that of ISV modelling in the low false alarm region for trials on NIST 2005 data. As the false alarm rate increases, a drop in performance is observed in NAP when compared to ISV modelling. This was not the case with NIST 2006 trials where both techniques maintained a consistent improvement over baseline results.

These results suggest that performing session variability compensation in the GMM domain provides no advantage over the implementation of the same technique in the SVM kernel space. It was also observed that both forms of session compensation produced larger performance gains for the NIST 2005 data than for 2006. It is hypothesised that this difference is due to the session variation of the projection training data being more representative of the conditions in the 2005 evaluation than for 2006.

5.1. Combined System and Score Fusion

A system, labelled as NAP ISV, was implemented combining GMMs trained using ISV modelling before SVM training using the NAP kernel. The projection matrix $U_n$ was retrained in this case on supervectors incorporating ISV modelling. The combined NAP ISV system attempts to remove any residual session variation from supervectors after ISV modelling has been applied. Figure 2 indicates that no significant performance gains are achieved through this combined approach over the ISV results for trials conducted on the NIST 2005 corpus. This trend was also observed in the NIST 2006 trials with results of the ISV and NAP systems being roughly equivalent to the NAP ISV configuration.

An alternative approach to a joint system is through the score-level fusion of the individual classifiers. For this task, linear fusion was implemented using the FoCal toolkit [14] to optimise linear log regression and minimise the mean-squared-error. The weights and bias for this fusion were estimated using the scores from trials conducted on the alternate corpora. That is, fusion parameters for NIST 2006 trials were obtained using scores from the NIST 2005 evaluation and vice versa. The technique of score fusion provided a small gain in performance in terms of minimum DCF for the NIST 2005 trials, however for the NIST 2006 trials, advantages were only observed in terms of EER.

The performance of the combined system and the fused score configuration indicates that the application of session compensation techniques in different domains is, to a large degree, not complementary. This is expected given the similar nature of the session variation modelling approaches as highlighted in Section 3.3.

6. Conclusions

This paper has compared the modelling of session effects during the generative modelling process as opposed to the discriminative modelling process in the context of a GMM mean supervector SVM speaker verification system. A comparison was made between incorporating inter-session variability modelling techniques during GMM training and nuisance attribute projection in the SVM kernel.

Evaluations on the NIST 2005 and NIST 2006 SRE conversational telephony speech corpora indicated that both ISV modelling and NAP gave very similar performance improvements over baseline results. The combination of the techniques into a single classifier — by both system level and score level fusion — provided only a small advantage.

7. References