Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Experiments in SVM-based Speaker Verification Using Short Utterances

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 investigates the effects of limited speech data in the context of speaker verification using the Gaussian mixture model (GMM) mean supervector support vector machine (SVM) classifier. This classifier provides state-of-the-art performance when sufficient speech is available, however, its robustness to the effects of limited speech resources has not yet been ascertained. Verification performance is analysed with regards to the duration of impostor utterances used for background, score normalisation and session compensation training cohorts. Results highlight the importance of matching the speech duration of utterances in these cohorts to the expected evaluation conditions. Performance was shown to be particularly sensitive to the utterance duration of examples in the background dataset. It was also found that the nuisance attribute projection (NAP) approach to session compensation often degrades performance when both training and testing data are limited. An analysis of the session and speaker variability in the mean supervector space provides some insight into the cause of this phenomenon.

1. Introduction

Considerable speech resources are typically used in the development of speaker verification technology leading to high levels of classification performance [1]. The practicality of such systems in the real world becomes questionable, however, when clients are required to provide lengthy utterances before access to a system will be granted. Reducing this requirement of sufficient speech while obtaining satisfactory performance has proved difficult as demonstrated in a number of recent studies [2, 3, 4, 5]. The adverse effects of limited speech intuitively has a large impact on forensics oriented applications in which the availability of sufficient and quality speech is not guaranteed. In light of this shortcoming in current technology, research continues to address the robustness of speaker verification technologies under such conditions.

In recent years, the Gaussian mixture model (GMM) mean supervector support vector machine (SVM) classifier has received considerable focus due to its successful application to the task of speaker verification [6]. Significant advances in the associated technology have resulted in the proposal of SVM kernels tailored to the speaker verification task and session variability modeling techniques [7, 8, 9]. As is common in the research field, these studies have focused on the NIST speaker recognition evaluation (SRE) corpora using training and testing utterances of approximately two and a half minutes of speech, from which good performance has been obtained. The question remains, however, as to the robustness of the GMM mean supervector SVM (GMM-Svec) classifier in the context of limited training and testing speech.

Motivation for an investigation into SVM-based speaker verification from short utterances is two-fold. Firstly, recent participation in the EVALITA 2009 speaker verification identity evaluations has highlighted the superior classification ability of the GMM-Svec classifier over joint factor analysis (JFA) GMM-based classification when ample speech is available, however, the opposite is true in the case of limited training and testing data [10]. The secondary motivation comes from recent studies into the effects of limited training data in the context of GMMs estimated using JFA [3, 4]. These studies demonstrated that session-compensation through JFA was more effective when the duration of speech used to estimate the speaker and session subspaces was matched to the evaluation conditions. Given the distinct information link between the GMM and SVM modeling domains in the GMM-Svec classifier [11], it is expected that similar attention should be placed on the data used in the implementation of session compensation techniques in the SVM kernel.

This paper analyses the effects of limited speech resources on the state-of-the-art GMM-Svec classifier in the context of text-independent speaker verification. The fundamental classifier components that are investigated in this study are briefly described in Section 2. Experimental results in Section 4 firstly illustrate the shortcoming of SVM-based verification of short utterances in comparison to the widely accepted GMM-based classifier. The effectiveness of each of the fundamental SVM system components is then analysed through a series of experiments. Focus is given to the duration of speech used in the SVM background, score normalisation and NAP transform training datasets. Highlighted in this study is the shortcoming of the common NAP approach to session compensation when limited training and testing speech is encountered. Subsequent analysis of this phenomena is also presented.

2. GMM-Svec Classifier Components

Discriminative modeling techniques are highly applicable to the task of speaker verification due to their inherent ability to distinguish a given client speaker from impostor speakers. Recent years have seen the GMM mean supervector SVM classifier [6] become one of the most widely adopted classifiers in the research community. Consequently, the GMM-Svec classifier regularly comprises part of submissions to the NIST speaker recognition evaluations (SRE) [1].

Maximising the performance obtained from the GMM-Svec classifier relies on the correct function of a number of fundamental components and techniques. This section outlines these system components and draws attention to the appropriate selection of utterances during system development.

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

2.1. Background Dataset

SVMs are trained to discriminate between positive and negative classes of training examples [12]. In the context of speaker verification, these are the client and impostor classes, respectively. The background dataset refers to the large collection of impostor examples used to discriminate against the client training examples in the speaker modeling process. Recent studies have highlighted the importance of selecting appropriate background examples to represent the evaluation conditions [13, 14]. These studies have also found impostor utterances of considerable duration to be particularly beneficial to the model training process and the subsequent performance achieved the system. The duration of speech used to train background mean supervectors is analysed in this study with regards to the amount of training and testing speech expected in the evaluation conditions. Mismatched training and testing durations are of particular interest where the impostor examples may be matched to either the short or long speech segment in a trial.

2.2. Session Compensation

Session variability compensation is an integral part of speaker verification technology in both GMM and SVM-based configurations and has been shown to significantly reduce classification errors [15, 16, 17]. Session compensation in SVM-based speaker verification is commonly employed using nuisance attribute projection (NAP) [7]. NAP attempts to counteract the adverse effects of session and channels variations by projecting the most dominant *nuisance* directions out of the SVM kernel space, thereby providing improved speaker discrimination. These directions are assumed to reside in a low-dimensional space are estimated from a held-out dataset containing multiple utterances from a large number of speakers. The estimation process involves decomposing the within-class scatter of this data and retaining the eigenvectors corresponding to the N highest eigenvalues in the matrix $\boldsymbol{U_n}$ (in this work $N = 40$). These directions can then be projected out of an input supervector, $\boldsymbol{m}$, using

$$\boldsymbol{m}_{\text{nap}} = \left(\boldsymbol{I} - \boldsymbol{U}_n \boldsymbol{U}_n^T\right) \boldsymbol{m} \tag{1}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{m}_{\text{nap}}$ represents the session compensated supervector.

Recent work regarding session variability modeling in the context of the JFA framework for GMMs has demonstrated that the benefits associated with session compensation rapidly decrease along with the duration of the test speech segment [3]. This is possibly due to the relatively high degree of within speaker variation attributed to high phonetic variation between these shorter utterances which is less dominant in longer speech segments. It seems apparent, therefore, to determine whether similar observations can be made in the context of NAP-compensated SVM-based speaker verification as speech resources become limited. Section 4.3 presents experimental results and a relevant discussion on the findings of these investigations.

2.3. Score Normalisation

Score normalisation techniques [18] are typically employed in speaker verification technology with the objective of counteracting statistical variations in classification scores. This is accomplished by scaling all scores to a global distribution where a client- and test-independent classification threshold can be applied. Z- and T-norm are commonly employed in combination to provide ZT-norm (that is, Z-norm followed by T-norm). Both techniques attempt to scale the output score distributions to have zero mean and unit variance based on the observed trends of an impostor score distribution. In the case of Z-norm, this impostor distribution is estimated by testing an impostor cohort of utterances against a given speaker model, whereas T-norm compares a given test utterance against a set of impostor speaker models trained from the cohort. Recent work has found little benefit from Z-norm in the context of GMM-based verification of short utterances [10], thus motivating an investigation into the observable benefits of score normalisation in the case of short utterance speaker verification using SVMs.

3. Experimental Configuration

The GMM mean supervector SVM system used in this study was previously described in [13]. GMM supervectors were produced through mean-only MAP adaptation using 24-dimensional, feature-warped MFCC features including appended delta coefficients. An adaptation relevance factor of $\tau = 8$ and 512-component models were used throughout. SVM training and classification was performed using 12288 dimensional GMM mean supervectors and the associated kernel [6]. The NIST 2004 SRE was used to form large gender-dependent background datasets. Examples from the background dataset were additionally used as the Z- and T-norm score normalisation cohorts as this configuration has been shown to perform well in [13]. Where applicable, NAP [7] was employed to remove the 40 dimensions of greatest session variability. Speech segments from the NIST 2004 SRE and Switchboard 2 copora were used to learn the nuisance directions.

The GMM-UBM configuration in Section 4.1 matched the system used to produce mean supervectors, however a relevance factor of $\tau = 32$ was used. Where applicable, JFA was employed using a 50-dimensional channel subspace and a speaker subspace of 200 dimensions. These subspaces were learned using the same dataset as specified for the NAP transforms. Where applicable, score normalisation was employed using the SRE'04 corpus as Z- and T-norm impostor cohorts.

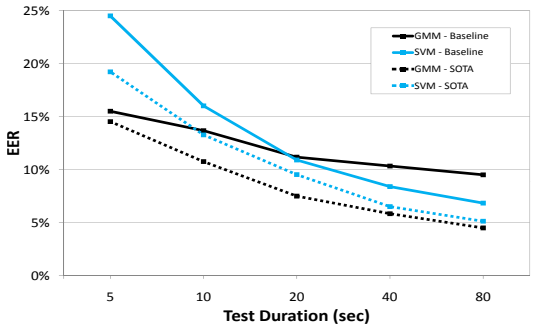
Evaluations in this work focus primarily on two cases of limited speech — (1) full training and limited testing data, and (2) limited training and testing data of equal duration. These conditions will be denoted *full-short* and *short-short*, respectively. Telephone-based utterances from the 1-sided, English-only condition of the NIST 2008 SRE were used for this task. These utterances were truncated to contain 5, 10, 20, 40 or 80 seconds of active speech (as determined using speech activity detector) from which GMM mean supervectors were trained. The first 5 seconds of active speech were removed from all truncated utterances to avoid potential overlap in the introductory speech.

4. Results

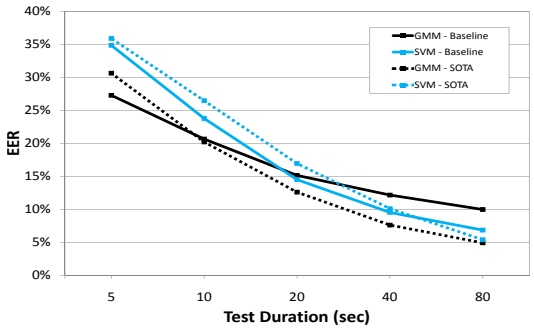
Following is an experimental study regarding the impact of limited speech on the fundamental components of the GMM-Svec configuration. These experiments look firstly at aspects of a baseline classification before progressively building towards a state-of-the-art configuration.

4.1. Baseline SVM Performance

Initial experiments were performed to determine the effects of limited speech on the GMM-Svec classifier that had been devel-



(a) full-short



(b) short-short

Figure 1: Trends in the GMM-UBM and GMM-Svec configurations for different durations of active speech in the (a) full-short and (b) short-short evaluation conditions.

oped toward the full-length *short2-short3* training and testing conditions of the SRE'08 and, therefore, does not specifically attempt to deal with the adverse effects of short speech segments. Performance statistics from the GMM-Svec SVM system were obtained in both baseline and state-of-the-art (SOTA) configurations, the latter of which incorporated session compensation and ZT-norm. Corresponding GMM-UBM configurations were also trialled to provide a point of reference from which to analyse SVM performance (see Section 3 for system specifications). The GMM-UBM configuration was selected for this purpose due to its stable operating characteristics under challenging evaluation conditions.

Figure 1 depicts the EER performance from baseline and SOTA SVM and GMM configurations for the *full-short* and *short-short* evaluation conditions on the SRE'08. The *full-short* trials in Figure 1(a) demonstrate that SVM performance degraded more rapidly than the GMM counterpart when the active test speech duration was reduced. This was particularly evident in the baseline systems (depicted in the plot as solid lines) in which the SVM performance provided significantly worse performance than the GMM configuration for short durations despite being superior when sufficient testing data was available. These observations can also drawn from the *short-short* results in Figure 1(b). Under these conditions, the SOTA SVM was found to offer worse performance than the baseline configuration when speech duration was restricted to less than 80 seconds while, in contrast, this was only observed in the SOTA GMM system when 5 seconds was used. The addition of session compensation and score normalisation to the baseline SVM configuration, in this case, resulted in reduced performance. Therefore, it would seem apparent that these common techniques must be tailored to deal with the conditions exhibited by short utterances in order to improve the robustness of SVM-based classification. The following sections aim to address this issue from a development data point of view.

4.2. Background Dataset

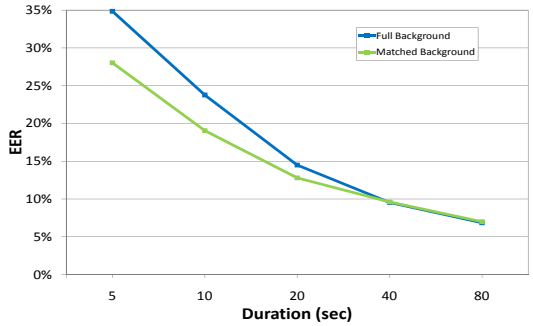
One of the fundamental differences between the GMM-UBM and GMM-Svec SVM classifiers is the use of an impostor or background dataset when training SVMs. While the background dataset may appear analogous to the world model (the UBM) in GMM classification, SVMs are not adapted from the background and instead, the SVM objective function actively seeks to discriminate the client training data from examples in the background. Previous studies have demonstrated the need to select appropriate background examples to match the evaluation conditions to provide good model quality [13]. This section investigates the amount of speech used in the training of the background supervectors in the context of limited training and testing conditions.

Due to the potential mismatch in enrolment and testing speech durations, several background dataset selection strategies were considered. These strategies included matching the duration of background utterances to either (1) the training duration, (2) the testing duration or (3) the duration of the shortest segment constituting a trial. For this task, a *short-full* condition was introduced in which the enrolment segment was truncated and the full-length test utterance was used for verification. Trial conditions were evaluated using an impostor dataset compiled from either full or short background utterances. Figure 2 depicts the EER from these trials over a range of test durations.

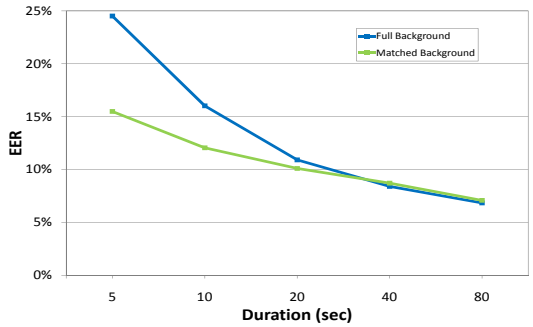
Figure 2 indicates that significant improvements tended to result from the matching of the background example duration to that of shortest segment in a trial. This, however, was not as evident in the case of the *short-full* trials of Figure 2(c), in which background matching was of no benefit when the training duration was above 10 seconds.

To provide clearer analysis, results specific to the evaluation conditions when using short segments of 10 seconds are detailed in Table 1. The *full-10sec* and *10sec-10sec* conditions in Table 1 exhibited significant performance gains when using background examples containing only 10 seconds of speech as opposed to full-length utterances. These relative improvements were up to 11% in minimum DCF and 25% in EER. The results from the last condition in the table, *10sec-full*, were inconclusive as to whether the background should be matched to the training, testing or shortest segment. However, when analysing these results along with the other evaluation conditions, certain consistencies were observed. Specifically, minimum DCF was improved when matching the background examples to the duration of the *test* segment, while the EER was minimised when matching to the *shortest* segment.

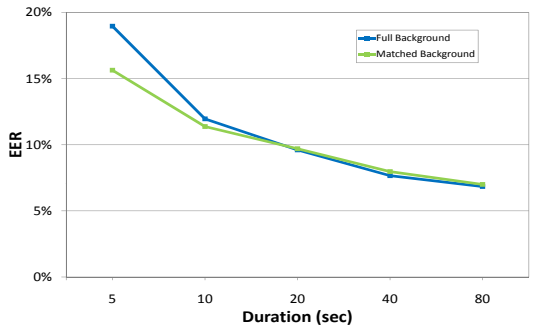
The observations drawn from the results in Table 1 indicate that matching the background example duration to that of the training segment does not always maximise performance. This finding is of interest when considering that the objective of the SVM training algorithm is to maximise discrimination between speaker and impostor classes. It would, therefore, seem intuitive to provide similar data for the classes being discriminated; in this instance, similar speech durations. It was demonstrated, however, that optimising discrimination against the characteristics of the data expected during verification or the most challenging data (i.e., shorter segments) resulted in a superior SVM client model.



(a) short-short



(b) full-short



(c) short-full

Figure 2: Comparing full and matched background selection strategies for the GMM-Svec configuration at different lengths of active speech for each evaluation condition.

The background strategy adopted for the remainder of this study is to match the duration of background utterances to that of the *shortest* utterance constituting a trial. It should be noted, however, that remaining experiments focus only on the *full-short* and *short-short* conditions in which the test utterance is also the shortest segment. This *matched* background configuration will be referred to as the Reference system for the purpose of analysing the effectiveness of NAP and score normalisation.

4.3. Session Compensation

Session compensation is an important component of speaker verification technology that typically improves classification performance by a considerable factor [17]. This section focuses on the application of session compensation using NAP in the context of limited speech. As mentioned in Section 2.2, NAP compensation relies on the appropriate estimation of a set of directions that best capture the observable session variability in the SVM kernel space from a transform training dataset. Experiments investigate the duration of utterances used to estimate this transform under two specific contexts — *full-short* and *short-short* evaluations.

Train-Test	Background	Min. DCF	EER
10sec-10sec	Full	.0815	23.77%
	10sec	**.0751**	**19.05%**
Full-10sec	Full	.0587	16.02%
	10sec	**.0520**	**12.05%**
10sec-Full	Full	**.0522**	11.96%
	10sec	.0560	**11.38%**

Table 1: GMM-Svec performance when using full and matched (10sec) background examples with 10 seconds of active training and/or testing speech.

4.3.1. Full-Short Evaluations

The previous section highlighted the importance of matching the duration of speech in background utterances to the testing or the shortest utterance in a trial. It is, therefore, hypothesised that examples in the NAP training dataset will exhibit a similar requirement in order the maximise the effectiveness of NAP in mismatching training and testing conditions.

The *full-short* trial condition was evaluated using NAP transforms estimated from full-length utterances and from utterances truncated to match the shorter, test utterance length. Figure 3 depicts the EER performance from these trials as a function of testing utterance duration along with the performance offered from the baseline configuration. For all durations trialled it can be seen that Matched NAP training consistently provided improved performance over the Reference and Full NAP training configurations. Comparable performance was, however, observed from the Matched NAP and Reference configurations when very limited data was available. The Full NAP results were particularly poor when less than 20 seconds of test speech was available such that the Reference configuration provided superior performance. In light of these observations, it is clear that matching the duration of utterances used to estimate the NAP transform to the shorter, test segment of a trial holds a distinct advantage over the estimation of the NAP transform from full length training utterances.

Section 4.2 demonstrated that the quality of SVM client models was improved when trained to discriminate the client training data against background examples representative of the test conditions. This observation is also apparent from the trials in Figure 3 such that compensating for the variations in the short test segment was found to be of greater importance than removing the variation observed in the enrolment utterance of sufficient length. Session compensation should, therefore, be targeted toward the variations observed in the speech segments from which the extraction of useful speaker information is more challenging. It should also be noted that the number of short background examples typically outweighs those of client training utterances by a considerable margin. Consequently, most of the discriminative information for SVM training is provided by the background dataset. It is apparent that reducing the interference of session variations on the informative impostor speaker characteristics in these examples also aids in the production of quality client SVMs.

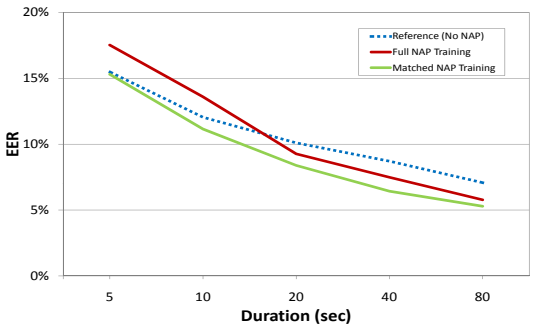


Figure 3: Comparison of NAP when estimating transforms from full or truncated utterances and evaluated on the *full-short* condition of the SRE'08.

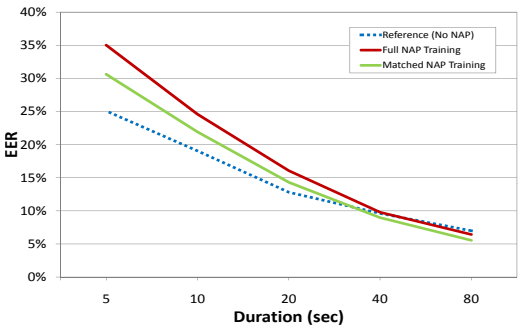


Figure 4: Comparison of NAP and Reference system performance when estimating NAP transforms from full or truncated utterances and evaluated on the *short-short* SRE'08 condition.

NAP Training	Min. DCF	EER
Baseline (No NAP)	**.0751**	**19.05%**
Full NAP training	.0838	24.59%
Matched NAP training	.0788	21.91%

Table 2: The use of NAP compensation in the 10sec-10sec condition when estimating the NAP transforms from full-length or 10 second utterances.

4.3.2. Short-Short Evaluations

The *full-short* evaluations demonstrated the need to match the NAP training data to the shorter test segment of a trial, however, limited gains over the baseline configuration were observed when very limited test speech was available. Of interest in the following trials is the effectiveness of NAP-based session compensation when both training and testing utterances are limited in duration.

Table 2 indicates the performance statistics obtained when applying session compensation using a NAP transform estimated from full or truncated utterances in the 10sec-10sec condition of the SRE'08 along with baseline system performance. It is clear from these results that the baseline system provides significantly better performance than either of the NAP compensated configurations. While the matching of NAP transform data to the limited speech conditions provided considerable improvements over a transform estimated from full-length utterances, its application to the baseline system degraded classification performance. This aligns with the findings of [4] in which the improvements expected of JFA-based session compensation were not observed when limited testing speech was encountered. In light of these observations, it would be beneficial to determine the amount of speech required by NAP in order for its application to benefit classification in the *short-short* conditions.

Figure 4 depicts the EER obtained when employing the full and matched NAP transforms along with the EER offered through baseline SVM classification. While considerable improvements were observed from NAP when 80 seconds of speech was available, the plot indicates that NAP struggles to provide any advantage over baseline performance when utterance duration was restricted below 40 seconds — even in the case of matched transform training utterances.

4.4. An Analysis of Session & Speaker Variability

The application of NAP to trials involving limited training and testing conditions degraded verification performance in Section 4.3.2. Analysis of the session and speaker variability observed in the SVM kernel space is expected to provide insight as to why NAP fails to benefit verification performance under these conditions.

In the context of JFA GMM-UBM speaker verification, previous studies have shown the observable variance in the session subspace to increase as utterance length is reduced [19]. It is believed that this increase in variance may to be due to the increased significance of phonetic variation between shorter utterances [3]. Given the distinct link between the GMM modeling domain and the SVM kernel space when using GMM mean supervectors, it is expected that the similar trends may be exhibited in the kernel space as available speech is reduced. In order to test this hypothesis, the magnitude of within and between scatter variance observed in the SVM kernel space was calculated to provide a measure of session and speaker variability, respectively.

These statistics were gathered from supervectors estimated using a MAP relevance adaptation factor of $\tau = 8$ (corresponding to the system configuration used throughout this study) and $\tau \approx 0$. In the case of $\tau = 8$, Table 3 details the magnitude of session and speaker variation observed in the SVM kernel space over a number of utterances lengths. It can be observed that the magnitude of session variation is reduced along with speech duration. This observation conflicts the findings of [19] and do not support the assumption that relatively high variation exists between short utterances. To investigate further these conflicting findings, the effect of relevance MAP adaptation on the SVM kernel space was was analysed. The statistics detailed in Table 3 were evaluated using a MAP relevance factor of $\tau \approx 0$ to essentially remove the influence of the UBM during supervector training. As expected, this allowed component means to move freely and provide an increase in variance magnitudes as observed in [19]. This draws attention to the significant influence of the relevance adaptation factor τ on the observable variations in the SVM kernel space.

Table 3 also indicates the ratio of session variance magnitude to speaker variance magnitude as observed in the SVM kernel space. It can be observed that the $\frac{\text{Session}}{\text{Speaker}}$ ratio is significantly greater for shorter durations of speech than for longer speech segments. Clearly, session variability is more dominant in the kernel space when using shorter speech segments causing

Duration	Session	Speaker	Session / Speaker
80 sec	0.473	0.358	1.32
40 sec	0.428	0.254	1.69
20 sec	0.334	0.154	2.17
10 sec	0.226	0.086	2.64
5 sec	0.137	0.045	3.06

Table 3: Magnitude of speaker and session variation observed in the SVM kernel space as utterance duration is reduced.

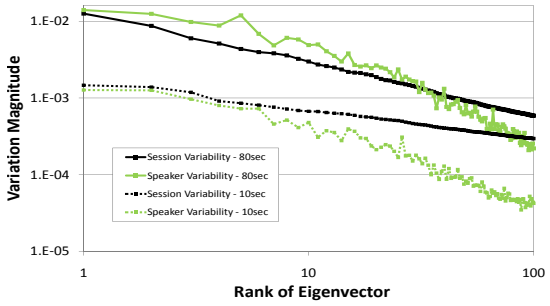


Figure 5: Session and speaker variability observed in the NAP transform training dataset comprising of 80 seconds and 10 seconds of active speech.

the verification task to become more challenging.

In order to gain an understanding as to why NAP is not effective when dealing with limited training and testing speech, the magnitude of session and speaker variation captured in the top 100 directions of greatest session variation[1] were plotted in Figure 5 when utterances of 80 and 10 seconds in duration were used to estimate the nuisance directions. Session variability is represented in this plot by the darker lines and speaker variability by the lighter lines. This plot shows that the slope of the session variability in the 80 second case is greater than that observed in training utterances containing only 10 seconds of speech while the slope of the speaker variability is similar in both cases. As highlighted in Section 2.2, NAP was developed based on the assumption that the vast majority of session variability could be expressed in a low-dimensional subspace. Figure 5, however, shows the slope of the eigenvalues to 'flatten' when reducing from 80 to 10 seconds of speech. This suggests that session variability becomes more isotropic as speech duration is reduced. Consequently, NAP fails to provide performance gains in these reduced speech scenarios because the assumption on which it was developed does not hold.

The development of techniques to address the issue of NAP-based session compensation highlighted in this section demands considerable attention. One such approach that may provide some improvement is scatter difference NAP (SD-NAP) [9]. The idea of SD-NAP is to introduce back into the NAP-compensated kernel space, a weighted influence of the between scatter statistics to ensure important speaker information is retained.

[1]The top 40 dimensions constitute the NAP transform used in this work.

Eval.	Cohort	Reference DCF	Reference EER	NAP (Matched) DCF	NAP (Matched) EER
Full-10sec	None	.0520	12.05%	.0470	11.15%
	Full	.0568	14.25%	.0496	12.30%
	Matched	**.0447**	**12.04%**	**.0461**	**11.07%**
10sec-10sec	None	.0751	19.05%	.0788	21.91%
	Full	.0750	19.46%	.0793	22.31%
	Matched	**.0749**	**18.81%**	**.0781**	**21.74%**

Table 4: The effect of matching ZT-norm score normalisation cohorts to limited speech evaluation conditions in Reference and NAP-compensated configurations.

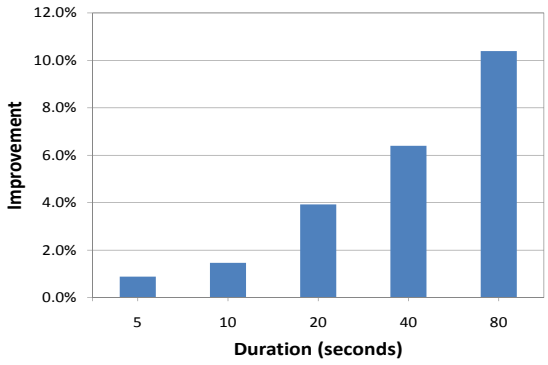


Figure 6: Relative minimum DCF improvements in the Reference configuration (No NAP) when applying matched score normalisation to raw scores.

4.5. Score Normalisation

As in the case of the SVM background dataset, suitable score normalisation cohorts much be selected in order to maximise the potential performance benefits [13]. This section briefly investigates how the duration of utterances in these cohorts corresponds to the effectiveness of score normalisation in the context of SVM-based classification with limited speech. While matching the score normalisation cohorts to the evaluation conditions is commonplace in systems submitted in the NIST SREs, the degree that this matching aids performance in the context of SVM-based speaker verification has not yet been reported in literature. To aid in discussion, results are presented only for speech durations of 10 seconds, however, it should be noted that similar observations were drawn from all other utterance durations. Table 4 presents the performance obtained when using Z- and T-norm impostor cohorts consisting of full-length utterances and utterance lengths matched to the evaluation conditions. The latter case corresponds to matching the T-norm cohort to the duration of the training utterance and the Z-norm utterances to the test duration of the evaluation protocol.

Results from the Full-10sec trials in Table 4 indicate a number of consistencies. Firstly, the full-length score normalisation cohorts provided the worst performance such that an increase in verification error was observed relative to the raw scores. In contrast, the best performance was obtained when matching the normalisation cohorts to the evaluation conditions. In this case, the Z-norm utterances consisted of only 10 seconds of speech while the T-norm segments remained full-length so as to match the client training conditions. In light of this observation, the selection of an appropriate Z-norm cohort alone had a significant

effect on performance under these limited speech conditions. While matched normalisation cohorts provided the best performance, the observable gains over the raw scores were limited with the exception of the 14% relative minimum DCF improvement in the Baseline system.

Similar to the Full-10sec evaluation conditions, Table 4 indicates that performance was maximised in the 10sec-10sec trials when truncating utterances in the score normalisation cohorts to 10 seconds. When comparing the scores between different normalisation cohorts and the raw scores, minimal variation can be observed. It should be noted based on the last row of Table 4 that score normalisation did not help to rectify the poor performance offered through NAP relative to the Reference configuration that was highlighted in Section 4.3.2. Figure 6 illustrates the relative improvements that were brought about by matched ZT-norm cohorts to raw scores of the Reference system for a range of speech durations in the *short-short* trial condition. Clearly, the benefits of score normalisation become less apparent as speech duration is reduced from 80 seconds down to 5 seconds. In light of these observations, the application of ZT-norm to SVM-based speaker verification with limited training and testing speech appears, to a large degree, to be unnecessary.

5. Conclusions

This paper presented a study on the effects of limited speech data on SVM-based speaker verification in the context of the GMM mean supervector SVM classifier. The fundamental components of this classifier were analysed when subject to limited training and testing data in the NIST 2008 SRE.

Initial experiments compared SVM-based classification performance to that of the widely accepted GMM-UBM configuration subsequently highlighting the relatively rapid degradation that SVMs exhibited as speech duration was reduced. The duration of utterances used to train the background dataset was found to have a considerable effect on classification performance. Matching these impostor utterances to either the shortest or the test utterance length expected in trials was found to significantly improve SVM-based performance.

NAP-based compensation was found to be most effective when estimating the nuisance directions from utterances containing an amount of speech matching the short, test speech segment of a trial. An issue with the common NAP approach was highlighted when both training and testing speech segments were limited to below 40 seconds such that degraded performance resulted from its application relative to baseline system performance. Finally, score normalisation was shown to be most effective when Z- and T-norm cohorts were matched to the evaluation conditions. However, it was found to provide few benefits when less than 20 seconds of speech was available. Based on these findings, it is apparent that future research should target the need for appropriate session compensation techniques in the context of SVM-based speaker verification using limited speech.

6. References

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

[2] M. McLaren, D. Matrouf, R. Vogt, and J .F. Bonastre, "Applying SVMs and weight-based factor analysis to unsupervised adaptation for speaker verification," *In print, Computer Speech & Language*, 2010.

[3] R. Vogt, J. Pelecanos, N. Scheffer, S. Kajarekar, and S. Sridharan, "Within-Session Variability Modelling for Factor Analysis Speaker Verification," in *Proc. Interspeech*, 2009, pp. 1563–1566.

[4] R.J. Vogt, C.J. Lustri, and S. Sridharan, "Factor analysis modelling for speaker verification with short utterances," in *Proc. IEEE Odyssey Workshop*. 2008, IEEE.

[5] M.W. Mak, R. Hsiao, and B. Mak, "A comparison of various adaptation methods for speaker verification with limited enrollment data," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, pp. 929–932.

[6] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[7] A. Solomonoff, W.M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2005, vol. 1, pp. 629–632.

[8] A.O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Ninth International Conference on Spoken Language Processing*, 2006, pp. 1471–1474.

[9] B. Baker, R. Vogt, M. McLaren, and S. Sridharan, "Scatter Difference NAP for SVM Speaker Recognition," in *Proc. International Conference on Biometrics*. 2009, pp. 464–473, Springer.

[10] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "QUT speaker identity verification system for EVALITA 2009," in Submitted to *Proc. International Conference on Information Sciences and Signal Processing and their Applications (ISSPA)*, 2010.

[11] 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.

[12] 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.

[13] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "Data-driven background dataset selection for SVM-based speaker verification," *In print, IEEE Trans. Audio, Speech and Language Processing*, 2010.

[14] M. McLaren, B. Baker, R. Vogt, and S. Sridharan, "Exploiting multiple feature sets in data-driven impostor dataset selection for speaker verification," in *To be presented in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010.

[15] R. Vogt and S. Sridharan, "Experiments in session variability modelling for speaker verification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2006, vol. 1, pp. 897–900.

[16] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[17] 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 International Conference on Acoustics, Speech and Signal Processing*, May 2006, vol. 1, pp. 97–100.

[18] 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.

[19] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Proc. Interspeech*, 2008, pp. 853–856.