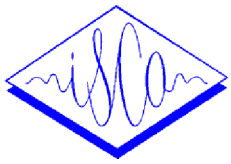
ISCA Archive
http://www.isca-speech.org/archive



Odyssey 2008: The Speaker and
Language Recognition Workshop
Stellenbosch, South Africa
January 21-24, 2008

Improving the performance of text-independent short duration SVM- and GMM-based speaker verification

Benoît Fauve[1], *Nicholas Evans*[2,1] *and John Mason*[1]

[1]Speech and Image Research Group, University of Wales Swansea, UK
[2]LIA, Université d'Avignon et des Pays de Vaucluse, France
{b.g.b.fauve.191992,j.s.d.mason}@swan.ac.uk
nicholas.evans@{univ-avignon.fr,eurecom.fr}

Abstract

In the task of automatic speaker verification (ASV) it is well known that the duration of the speech signals is an important factor in the ultimate accuracy of the system. This paper deals with some of the aspects of adapting systems to work with limited amounts of data. First we highlight the importance of a well-tuned speech detection front-end when working with short durations. We consider a well-established technique (GMM) as well as a recent development (SVM on GMM mean supervectors), showing their limitations and alternatives. In particular the benefit of eigenvoice modelling in the context of short duration tasks is highlighted. Finally experiments on standard NIST databases demonstrate fusion potential between the presented techniques and significant gains when compared to a single GMM.

1. Introduction

Interest in text-independent automatic speaker verification (ASV) has grown significantly over the last several years as evidenced by the annual speaker recognition evaluations (SREs) administered by the National Institute of Standards and Technology (NIST) [1]. They currently offer one of the most viable means by which researchers can compare and contrast different approaches on common and meaningfully sized databases. For each SRE there are a number of different conditions spanning a range of data quantities for training and testing scenarios. Since the NIST SRE'04 the one condition in which participants are required to participate relates to one side of a five-minute-long telephone conversation. This gives approximately 2.5 minutes of speech per person and with a second such conversation 2.5 minutes are available for both training and testing.

Given that this is a compulsory condition for NIST entry, inevitably this condition has received by far the greatest attention in the literature. In particular some of the latest developments in the field which tackle the well known problem of intersession variability namely factor analysis (FA) [2, 3] and nuisance attribute projection (NAP) [4,5] have led to meaningful improvements in performance on tasks involving a few minutes of speech or more, with error rates having roughly halved over the last three campaigns [6].

However, commercial and practical situations call for speaker verification using much shorter speech durations and, perhaps surprisingly, these conditions have received comparatively little attention within the research community. For example, two sites who report state-of-the-art performance on the NIST SRE'06 database [7, 8] do not report any results on the shortest duration tasks involving just 10 seconds each for both training and testing.

Aside from the obvious difficulties associated with training or adapting accurate or reliable speaker models using such limited data quantities one possible explanation for this observation, in addition to that stemming from NIST's focus on the longer duration tasks, may be attributed to the fact that, as we have shown in our recent paper [9], some recently proposed channel compensation techniques are not easily transposable to the shorter duration tasks. In [9] we show that a system based on Gaussian mixture models (GMMs) and whose front-end is optimised for long duration tasks is very much suboptimal when applied to the shorter duration tasks. We extend this recently published work by illustrating in this paper a sensitivity to speech activity detection (SAD), a problem only observed when dealing with limited amounts of speech. We subsequently highlight the limits of the traditional maximum a posteriori (MAP) model adaptation approach for the short duration scenario.

An original contribution in this paper is the application of eigenvoice (EV) principles with improved performance in the context of short duration task (10s10s). Improvements are particularly good when scores are fused with those of a GMM-MAP system. We then propose a series of changes for a support vector machine-based system with a GMM supervector linear kernel (SVM-GSL) [10]. Finally results and fusion potential are validated on the latest NIST'06 database.

Nicholas Evans is now with Institut Eurecom, Sophia Antipolis, France

The remainder of the paper is organised as follows. In Section 2 we introduce the protocol and systems used throughout the paper. Section 3 deals with sensitivity to SAD. Limits and alternatives of ASV systems are discussed in Section 4 for GMMs and in Section 5 for SVM-GSL. In Section 6 we present fusion and validation results. Our conclusions are drawn in Section 7.

2. Protocol and systems

The NIST SRE'04, SRE'05 and SRE'06 databases are used for all experimental work reported in this paper. Of the different durations in the three databases we focus on two durations specifically. They are: (i) 1conv4w, an average of 2.5 minutes of speech and (ii) 10sec4w with an average of 10 seconds of speech. We refer to these two durations from now on as 1c and 10s respectively.

We first consider the 1c1c task ($\sim$2.5 minutes of speech for both training and testing), the 1c10s task ($\sim$2.5 minutes of speech for training but with now only $\sim$10 seconds of speech for testing), the complement condition 10s1c, and finally 10s10s. In the second part of the paper we concentrate only on the 10s10s condition.

For all experiments reported in the paper the background data are as defined in our previous publication [6] and all come from the NIST'04 database. For the system optimisation stage we conduct development experiments on the male part of the NIST'05 database and protocols, leaving the entire NIST'06 database, male and female and all languages, for final validation only. In all cases performances are assessed with the minimum of the decision cost function (minDCF) in accordance with NIST's definition and in terms of equal error rates (EER).

The systems presented are developed using SPro[1] and ALIZE[2] which are both open source toolkits. Full descriptions of their use in this work and for NIST SREs generally are available in [6, 11]. Note that in all our systems involving GMMs, model warping [11] is used.

3. Sensitivity to speech activity detection

One of the first stages of processing in ASV is to determine intervals of speech along the time course. Simple, effective and popular approaches are based on energy distributions. Here we consider a form proposed in [11].

3.1. Mean and Weight based approaches

We consider two model based speech activity detection (SAD) variants. For both, a tri-Gaussian model is fitted to the energy component of a speech sample. An energy threshold is used to distinguish speech from non-speech and only those frames whose energy component is above the threshold are retained for speaker modelling and testing. The threshold is determined according to one of two characteristics corresponding to the highest energy Gaussian. A weight-based SAD (wSAD) selects a threshold according to the weight of the highest energy Gaussian component, whereas a mean-based SAD (mSAD) selects its threshold from the mean and standard deviation. Further details of both approaches can be found in [11].

In both cases a parameter α_{SAD} may be used to tune the SAD varying the number of frames which remain after threshold application. In [9] we show the influence of this parameter with mSAD. Our results suggest that the highly selective value of α_{SAD} derived to give state-of-the-art performance on the 1c1c task proves to be overly selective when applied to the shorter duration task. In this paper, we present new experiments which compare the two different approaches with different speaker verification systems and different task duration conditions.

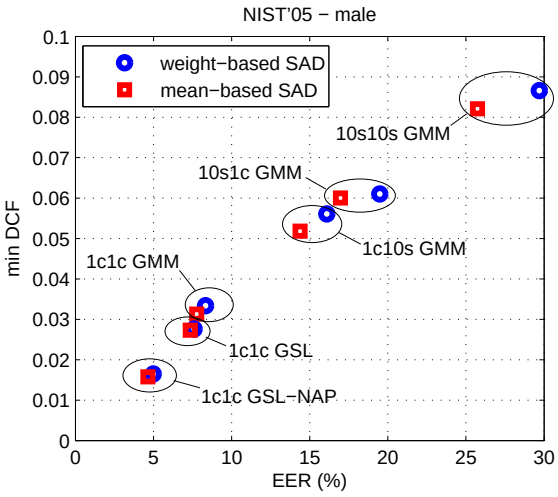


Figure 1: *Graph of MinDCF against EER showing ASV performance for range of data durations, 1c1c, 1c10s, 10s1c and 10s10s for a standard GMM configuration on the male part of the NIST'05 database. The points for SVM-GSL and SVM-GSL-NAP are also included for further comparison on 1c1c.*

3.2. Performance

Figure 1 shows performance in terms of minDCF against EER for a basic GMM system for 1c1c, 1c10s, 10s1c and 10s10s task and with SVM-GSL [10], SVM-GSL-NAP [5] on the 1c1c task (SVM-GSL and SVM-GSL-NAP results are not shown for the shorter durations as they are suboptimal, a point we discuss further in Section 5). These results come from numerous optimisation experiments varying the SAD threshold parameter. All tests relate to the male part of the NIST'05 SRE protocol.

In each case two points are illustrated, one for weight-based SAD (wSAD) and one for mean-based SAD (mSAD). In all cases the points relate to individually opti-

[1] http://gforge.inria.fr/projects/spro

[2] http://www.lia.univ-avignon.fr/heberges/ALIZE/

mised settings of α_{SAD}. There are two main observations we can draw:

- wSAD is clearly suboptimal for the 3 short duration tasks (10s10s, 10s1c and 1c10s). We conducted an extensive set of experiments (not reported here) where the frame selection sensitivity was varied according to the α_{SAD} parameter of wSAD for the 10s10s task. Even when GMM parameters (the number of Gaussians and relevance factor in MAP adaptation) were optimised, in all cases the performance fell short of that obtained with mSAD.
- the difference in performance between the wSAD and mSAD approaches diminishes for the longer duration tasks for the basic GMM and also for the two more recent approaches (SVM-GSL and SVM-GSL-NAP).

This observation confirms the idea we highlighted in [9], namely that GMM systems that exhibit state-of-the-art performance on longer duration tasks can have suboptimal performance when applied to shorter duration tasks. In this case a system optimised with weight-based SAD on longer durations may prove to be suboptimal when applied to shorter durations such as 10s10s.

An explanation may be found in the quality of the SAD. Table 1 shows the mean and standard deviation of the amount of speech, in seconds, which remain post-SAD for 4 different SAD configurations. α_{SAD} values are chosen for illustration purposes, to present examples for mSAD and wSAD providing a similar average of frames ($\sim$110s as with ASR transcripts and $\sim$85s as in [6]) in two quite different cases. These numbers are obtained from 219 1c, male samples from the NIST'04 database.

Table 1: *Statistics (average and standard deviation) of the amount of speech in seconds from 219 1c, male samples from the NIST'04 database according to ASR transcripts (line 2), and amount found for 4 configurations of SAD (lines 3-6).*

	Mean (s)	Std (s)	Ratio
ASR transcripts	**110.9**	38.9	2.9
mSAD, $\alpha_{SAD} = 0$	**112.4**	27.0	4.2
mSAD, $\alpha_{SAD} = -0.5$	83.2	20.1	4.1
wSAD, $\alpha_{SAD} = 0.3$	**110.8**	24.9	4.5
wSAD, $\alpha_{SAD} = 0$	86.8	15.5	5.6

Also illustrated (row 2) are similar statistics derived from the ASR transcripts which are provided with the NIST database. They are used here as a baseline with which to compare the SAD derived values. The ASR transcripts suggest that there is an average of 110.9 seconds of speech per 1c file and a standard deviation of about 39 seconds, the ratio between the mean and standard deviation being 2.9. Turning now to the SAD derived statistics we observe mSAD ratios of 4.2 and 4.1 and wSAD ratios of 4.5 and 5.6. We notice that both SAD result in a higher ratio corresponding to a lower variance in the number of frames found over the 219 1c files. Whereas that ratio stays constant in the case of mSAD, it varies significantly for wSAD. This limited evidence suggests that mSAD better reflects (though not perfectly) the actual amount of speech available than wSAD does.

Taken together with the results presented in Figure 1 it therefore seems that the accuracy of the SAD is especially important for short duration tasks. For longer durations the various averaging in the GMM modelling and scoring attenuates the effects of a less well performing SAD.

4. GMM adaptation and short durations

After this front-end consideration we now concentrate on ASV systems, starting with the well-established GMM systems.

4.1. Limitation of MAP approach

One of the key elements involves the maximum a posteriori (MAP) adaptation of a UBM. A speaker-specific, client model is adapted from the UBM using observed data X by modifying the GMM mean parameters according to:

$$\mathbf{m}_i(X) = \alpha_i \mathbf{E}_i(X) + (1 - \alpha_i)\mathbf{m}_{wi} \tag{1}$$

and

$$\alpha_i(X) = \frac{n_i(X)}{n_i(X) + r} \tag{2}$$

where n_i is the posterior probability, $\mathbf{E}_i$ is the expected value of the observed data and $\mathbf{m}_{wi}$ and $\mathbf{m}_i(X)$ are respectively the UBM and model GMM mean vectors, all corresponding to the i^{th} Gaussian. r is the relevance factor which acts to control the degree of adaptation as per [12].

Simply stated, n_i corresponds to the number of frames close to the i^{th} component. For the shortest duration tasks only a few Gaussian components will have close enough frames to be significantly and accurately adapted and overall $m_i X$ might be a poor estimation of the client GMM. To illustrate this point we report an experiment where Equation 2 is not used and is instead replaced by a constant adaptation coefficient (CT) namely $\alpha_i(X) = \alpha(X) = \alpha$. Results with this adaptation technique are reported in table 2 and compared to MAP. In this case we obtained poorer performance on the 1c1c

task but similar or slightly better performance when using a constant α set to 0.3 (CT in table 2) on the 10s10s task. When approaching the shortest duration tasks, i.e. with a diminishing number of available frames (for example an average of only 595 frames are used for model training on the 10s10s task for NIST'05 database, male part in our optimal configuration), we reach a potential limit of the MAP adaptation algorithm.

Table 2: *Performance comparison of MAP adaptation technique to a uniform constant adaptation (CT) on development set NIST'05 male only.*

Task	1c1c		10s10s	
Adaptation	MAP	CT	MAP	CT
EER(%)	8.25	9.07	25.6	24.7
minDCF(x100)	3.18	3.34	8.09	8.06

4.2. Assessment of eigenvoice modelling

An interesting technique that could deal with such extreme conditions is eigenvoice (EV) modelling [13]. The main idea behind this approach and its use for speaker verification [14–16] is that the mean supervectors, $\mathbf{m}$, of the client model are constrained to follow:

$$\mathbf{m} = \mathbf{m}_w + \mathbf{V}\mathbf{x}, \tag{3}$$

where $\mathbf{V}$ is a low rank matrix (of rank K), base of the eigenvoice space. As $K \ll C \times F$ (C being the number of components in the GMM and F the feature order) the number of free parameters is drastically decreased; this helps in parameter estimation. As this technique seems especially well suited for conditions with limited amounts of data we now report an experiment to assess its potential. In this experiment the mean supervectors of the client models are calculated as follows:

$$\mathbf{m} = \mathbf{m}_w + \mathbf{V}\mathbf{V}^T(\mathbf{m}_{MAP} - \mathbf{m}_w). \tag{4}$$

$\mathbf{m}_{MAP}$ is the speaker mean supervector derived from MAP adaptation (Equation 1). $\mathbf{V}$ is the base of the eigenvoice space, derived via principal components analysis (PCA) on mean supervectors from a set of well-trained speakers. It lies where differences between speaker models are found to be the most pronounced. To calculate the speaker models, we project the mean supervectors onto the eigenvoice subspace after MAP adaptation. It can be viewed as a post process of the MAP adaptation that removes poorly adapted dimensions. Our experiments using $\mathbf{E}(X)$ (unadapted mean, 'Maximum Likelihood' version of the parameter estimation) instead of $\mathbf{m}_{MAP}$ led to poorer performance. The obtained mean parameters are used for the speaker model. The scoring is the same as for a traditional UBM-GMM approach.

Table 3 shows some results of such a technique compared to the traditional MAP adaptation on both 1c1c and 10s10s task. Such a comparison has already been reported in for example [16] but without distinguishing between task durations. By considering separately tasks by their duration, we see that if eigenvoice modelling (GMMev) proves to be suboptimal on the 1c1c task, results in Table 3 highlight the benefit of eigenvoice modelling when working with sparse training data.

Table 3: *Performance in terms of EER for MAP and eigenvoice modelling (GMMev) on 1c1c and 10s10s task from development set NIST'05 male only.*

EER(%)	1c1c	10s10s
GMMev	12.51	23.65
GMM-MAP	8.69	25.52

More details on the system setup and further results including combinations with other systems are discussed in Section 6.

5. SVM-GSL and short duration tasks

In Section 3 we present some results with SVM-GSL and SVM-GSL-NAP on long duration tasks only. The reason is that the performance of such systems are suboptimal when compared to that of GMM when the available data is limited.

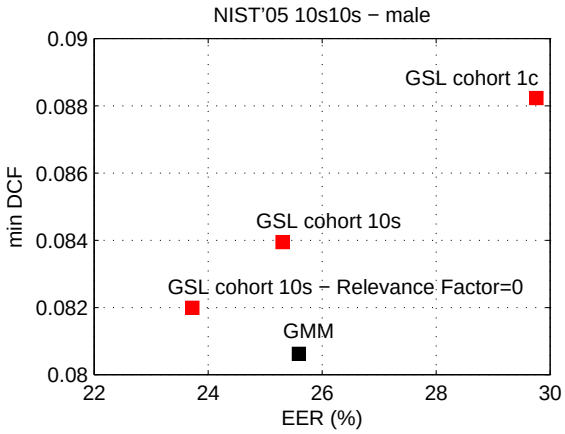


Figure 2: *MinDCF against EER for GMM baseline and a series of GSL systems on development set NIST'05 male only.*

This is illustrated on Figure 2, where a system as used in [6] and referred to 'GSL cohort 1c' is clearly outperformed by a GMM system in terms of both minDCF and EER. The NAP compensated version of such an SVM system [4] (not plotted here) proved even more suboptimal (0.0998 for minDCF and 34.4% for EER). Our effort to utilise NAP on short durations have been unsuc-

cessful; however, in the following paragraphs we present some improvements that can be made to the SVM-GSL system.

A first modification is to match the duration of the negative cohort examples with that of the data for the model being trained ('GSL cohort 10s' in Figure 2). Such a matching is also used in score normalisation techniques and referred to for example as 'targeted T-norm'. For SVM-based speaker verification 'targeted cohorts' seem to be beneficial.

Another change that brings further improvements is to directly use $\mathbf{E}(X)$ (the adaptation coefficient α_i in Equation 1 is set to 1), in the speaker models. On Figure 2 this system is marked 'GSL cohort 10s Relevance Factor=0' as with $\alpha_i = 1$ correspond to a null relevance factor in MAP adaptation.

Model warping [11] is still applied. Model warping assures that each feature dimension follows a global 0-mean and unity variance distribution. This is a priori information on the way the model should describe the acoustic distribution. When applied on unadapted GMM means (α_i=1 in equation 1), it leads to better performance with a SVM-GSL system on 10s10s. Note that such a technique is suboptimal for longer durations or when applied in a traditional GMM framework. Also of interest, using eigenvoice modelling as presented in Section 4 to obtain the GMM supervector brings poorer performance when used with an SVM classifier (results not presented here).

The difficulties encountered in Section 4 and 5 to find useful information to be exploited by GMM or SVM systems show the problem, when confronted with short durations, to find reliable statistics from the limited amount of available speech.

The series of modifications presented above brings an overall relative improvement of 7.0% on the minDCF and 20.5% on the EER.

6. Results

In this section we present individual and combined results on the 10s10s task from the previously described systems.

6.1. GMM, GMMev, GSLw and fusion

3 systems are presented here:

- GMM: our GMM baseline with 33 feature coefficients (16 LFCC, 16 Δ and Δ energy) with mSAD and α_{SAD}=0. The same front-end is used for all other systems to ensure fusion results come from system complementarity and not a front-end variation.

- GMMev: as described in Section 4. 124 males speakers and 185 female speakers from NIST'04 are used to derive the gender dependent eigenvoice spaces. The rank of the matrix $\mathbf{V}$ is K=100 for male and K=140 for female (more female individual being available in NIST'04 database).

- GSLw: SVM-based system as described in Section 5. The 'w' tag highlights the importance of model warping in this approach.

Fusion is performed from an unweighed sum of T-normalised scores. Our attempts to use a more sophisticated fusion approach (logistic regression) did not bring any meaningful improvement. An observation is that the results on the development set translate well to the validation set, namely the NIST'06 both genders, all languages ('DET1' as referred to by NIST).

Performances of the 3 single systems are shown in Figure 3 and 4. The fusion of two or all systems further demonstrates complementarity between the different approaches with significant improvements.

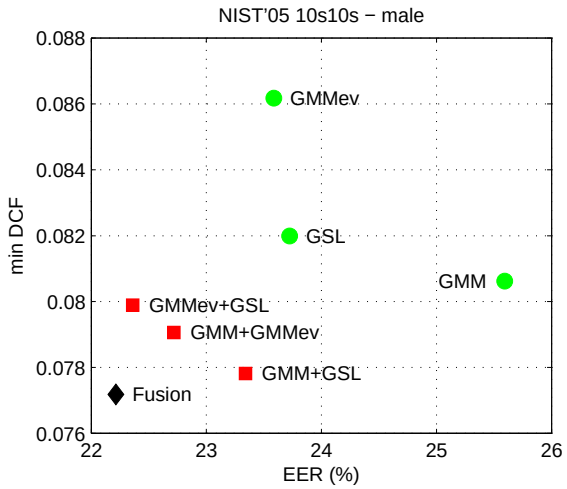


Figure 3: *MinDCF against EER for GMM, GMMev and GSL systems and their combinations on development set NIST'05 male only.*

Finally we also show fusion results with our submission to the NIST'06 SRE for the 10s10s task. UWSsub is a linear fusion between 3 GMM systems using different front-ends. It is interesting to see in Figure 4 that GMMev and the optimised SVM-GSL show similar complementarity with this system as with the single GMM. This final result shows a relative improvement of 7.1% on the minDCF and 16.1% on the EER when compared to GMM mSAD (whose front-end SAD has been optimised).

6.2. DET curves

Figure 5 shows 3 DET plots. The first is from our (UWS) NIST'06 1c1c submission. It is a standard GMM system and the plot here represents a typical example of performance on the short duration task when no specific optimisation to the length of the task is done. The middle curve

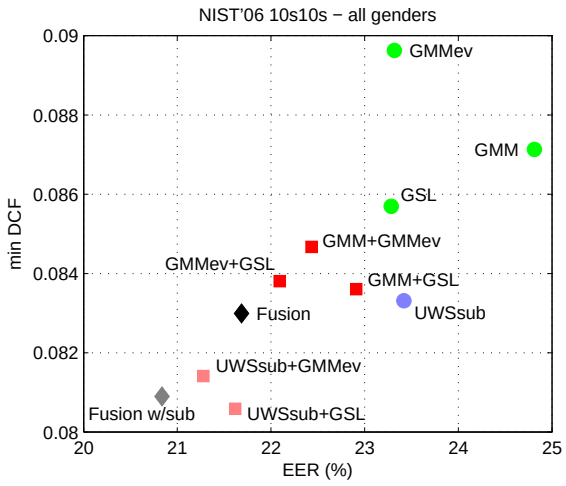


Figure 4: *MinDCF against EER for GMM, GMMev and GSL systems and their combinations on validation set NIST'06 both genders all languages (DET1). Results with UWS submission to NIST'06 are also given.*

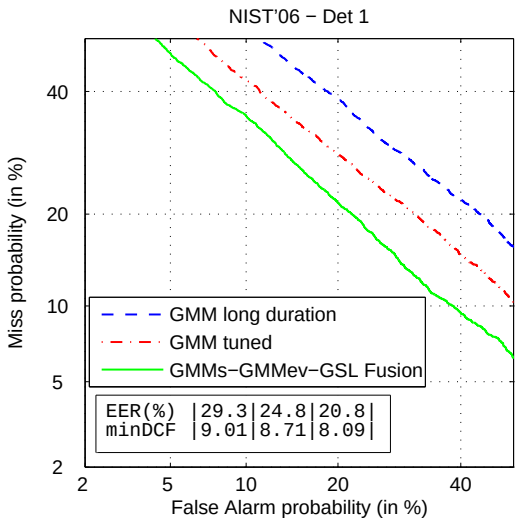


Figure 5: *DET curves for a standard GMM optimised on long duration, a similar GMM but tuned for short durations and finally a combination of GMMs, eigenvoice modelling GMM and SVM-GSLw approaches; all on NIST'06 10s10s task.*

comes from a GMM-MAP system with tuned parameters (frame selection, feature dimension, number of Gaussian components) as reported in this paper and in [9]. Finally we present the curve for the fusion results as described in Section 6.1. This result includes two key elements for short duration:

- new approaches beyond the traditional GMM-MAP and
- fusion between many systems.

As the amount of speech data becomes sparse the combination of expert systems proves particularly beneficial.

7. Conclusions

This paper highlights the limitations of transposing state-of-the-art techniques that have been used widely on longer duration tasks to shorter duration tasks. We first show the importance of accurate speech detection. Our experiments demonstrate the high sensitivity to SAD parameters with short duration tasks. We then show the limits of both GMM and SVM-GSL cases with MAP adapted mean parameters and propose some novel alternative solutions. Finally the benefit of eigenvoice modelling on the short duration task is highlighted. Meaningful improvements are demonstrated on the standard NIST databases; however, further work is needed to better understand and integrate all potential variabilities in ASV, amounts of speech being an important one.

8. References

[1] A. Martin and M. Przybocki, "The NIST speaker recognition evaluation series, National Institute of Standards and Technology's website, http://www.nist.gov/speech/tests/spk."

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in gmm-based speaker verification," *Audio, Speech and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1448–1460, 2007.

[3] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independant speaker verification," in *Proc. Interspeech*, 2005, pp. 3317–3320.

[4] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP*, vol. 1, 2005, pp. 629–632.

[5] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *ICASSP*, 2006.

[6] B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. Mason, "State-of-the-art performance in text-independent speaker verification through open-source software," *Audio, Speech and Language Processing, IEEE Transactions on*, vol. 15, Issue 7, pp. 1960–1968, 2007.

[7] R. Matejka, L. Burget, P. Schwarz, O. Glembek, K. M., F. Grezl, J. Cernocky, D. van Leeuwen, N. Brummer, and A. Strasheim, "STBU system for

the NIST 2006 speaker recognition evaluation," in *Proc. ICASSP*, vol. 4, 2007, pp. 221–224.

[8] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Politecnico di Torino 2006 NIST speaker recognition evaluation system," in *Proc. Interspeech*, 2007, pp. 1238–1241.

[9] B. Fauve, N. Evans, N. Pearson, J.-F. Bonastre, and J. Mason, "Influence of task duration in text-independent speaker verification," in *Proc. Interspeech*, 2007, pp. 794–797.

[10] W. M. Campbell, D. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 210–229, 2006.

[11] J.-F. Bonastre, N. Scheffer, C. Fredouille, and D. Matrouf, "NIST'04 speaker recognition evaluation campaign: new LIA speaker detection platform based on ALIZE toolkit," June 2004, NIST SRE'04 Workshop: speaker detection evaluation campaign. Toledo, Spain.

[12] D. A. Reynolds, T. Quatieri, and R. Dunn, "Speaker recognition using adapted mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[13] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proc. ICSLP*, vol. 5, 1998, pp. 1771–1774.

[14] S. Lucey and T. Chen, "Improved speaker verification through probabilistic subspace adaptation," in *Proc. Eurospeech*, 2003, pp. 2021–2024.

[15] P. Kenny and P. Demouchel, "Eigenvoices modeling with sparse training data," *Audio, Speech and Language Processing, IEEE Transactions on*, vol. 13, pp. 345–354, 2005.

[16] J. Mariéthoz and S. Bengio, "A comparative study of adaptation methods for speaker verification," in *Proc. ICSLP*, September 2002, pp. 581–584.