Speaker adaptation in the NIST Speaker Recognition Evaluation 2004.

David A. van Leeuwen

TNO Human Factors,
Soesterberg, The Netherlands.
david.vanleeuwen@tno.nl

Abstract

New in the 2004 edition of the NIST Speaker Recognition Evaluation (SRE) was the condition where unsupervised adaptation of speaker models is allowed. Despite the promising results on development test material, hardly any beneficial results were obtained in the Evaluation itself. An analysis is made why this was the case, and it appears that a mimimum level of performance is essential to obtain results using adaptation that improve on the performance without adaptation. Further, the system should be well calibrated. For the conditions with 8 conversation sides we have been able to find improvement using unsupervised adaptation using the NIST 2004 evaluation, both for an UBM/GMM adaptation methodology, and a novel SVM adaptation methodology. The minimum DCF for a fused system drops from 0.259 for the unadapted condition to 0.231 for the adapted condition.

1. Introduction

Adaptation in Automatic Speaker Recognition has recently become a subject of study in literature [1, 2]. The basic idea is, that after a short enrollment in a speaker verification system, normal operational use of the system will lead to many *true speaker* trials and will thus provide the system with more of the target speaker's speech material that can be used to improve the target speaker model. An obvious caveat of this technique is that a false acceptance trial will deteriorate the target speaker model and will therefore have a negative impact on the detection performance. The correct choice of a decision threshold (calibration) during adaptation is therefore not only important for a decision evaluation measure such as the *actual detection cost* C_{DET} [3], but of direct consequence to a score evaluating performance measure such as the DET curve [4] as a whole. In the context of the yearly NIST text-independent Speaker Recognition Evaluations (SRE) [3] unsupervised adaptation was proposed by Claude Barras at the SRE 2003 workshop, and subsequently introduced as an optional evaluation condition in the NIST SRE 2004.

As noticed in [2], the evaluation priors (the fraction of the number of non-target and target trials) play an important role in the final performance of a speaker adapting system. This is unlike the influence on traditional performance measures such as the position of the DET curve (score evaluation) or the value of C_{DET} (decision evaluation). In the latter measure externally determined *application priors* are parameters of the cost function, often set by the evaluator [5] Notice the difference in the evaluation priors between a speaker verification system in a real-world password-based access control system, where impostor trials are difficult to find [6], and in the text-independent NIST speaker recognition evaluation, where target trials typically form only 10 % of all trials.

In this paper we describe the development of the TNO 2004 speaker adaptive speaker recognition system and its performance in the NIST SRE 2004. Then we address the differences in performance on the development test database and the real evaluation. As a result of that, we have adapted the training protocol and applied some improvements in the Gaussian Mixture Model (GMM) based speaker recognition system. Further, a speaker adaptive methodology for the TNO Support Vector Machine (SVM) based system is introduced here as a novelty, and the results of both the improved GMM system and the new SVM adaptive system are finally evaluated on part of the NIST SRE 2004 data set. We intend to submit this speaker adaptive system to the NIST 2005 SRE, of which the results will be available at the Interspeech conference.

2. Experimental setup

2.1. Evaluation speech data bases

Two evaluation databases were used in this paper. The first is the database used for the one-speaker limited data detection task of the NIST SRE 2002. This database consists of 191 female and 139 male speakers. The data is recorded using the 'Switchboard' methodology and consists of excerpts of cellular telephone conversations. For model training typically 2 minutes of speech material is available, while test segments have durations of about 3–60 seconds. All speech is primarily in the English language.

The second evaluation database used is the NIST SRE 2004 data set. Of the many conditions evaluated there [3], we will focus on the '1side-1side' and '8sides-1side' conditions. In these conditions the test segments contained a whole speaker conversation side, and the model training material consisted of either one conversation side (of approximately 5 minutes) or 8 conversation sides, respectively. The data is part of the 'MIXER' data collection effort, and contains many independent variables that are of interest to speaker recognition research. The evaluation contains different languages and includes many trials for which the training and testing material consists of different languages. Also, there is a great variation in handset and channel type within the database.

2.1.1. Speaker recognition systems

Two speaker recognition systems are used in this paper. These are primarily the 'snapshot' versions of the ongoing development of the TNO system, taken at the NIST SRE 2004 and

2005. The TNO 2004 system consists of up to 5 different subsystems, all of which operate on acoustic features. One of them is a system based on Support Vector Machines [7], the other are varieties on the popular Universal Background Model (UBM) technique [8]. The four UBM/GMM systems vary in number of Gaussians and UBM model training, and in acoustic features. For this article we will discuss the two best performing GMM subsystems.

The 2004 GMM subsystems used as features 13 PLP coefficients plus deltas, sampled every 16 ms. Silence is removed by discarding all frames with an energy less than 30 dB below the utterance maximum. Features are then normalized by feature warping [9]. Subsystem 1 consisted of a 512 mixture UBM, trained on NIST SRE 2001 training speakers, and used relevance factor $r_m = 16$ for MAP adaptation [8] of speaker models. Subsystem 2 consisted of a 1024 mixtures UBM trained on 324 male and 256 female speakers of the Switchboard 2 phase 2 release. Because the channel condition were quite different from the SRE 2002 development test material, we used a quite 'aggressive' value $r_m = 6$. Gender specific UBM's and T-norming models were applied.

The TNO 2005 system contains two adapting acoustic subsystems, one GMM and one SVM. This year we have implemented the 'feature mapping' approach [10]. We use the same PLP unmapped features as in the 2004 system, except that we use utterance-based zero mean, unit variance normalization. A 2048-component 'root UBM' has been trained on 591 speakers of Switchboard 2 phase 2 landline data and NIST SRE 2001–2003 cellular data. For four acoustic/channel conditions and two speaker genders we adapted 8 channel specific GMM's from this root UBM, adapting only Gaussian means. The acoustic channels were based on automatic Carbon/Electret landline microphone classification by MIT-LL and GSM and CDMA cellular coding channels as provided by NIST evaluation key databases. Both in training and testing, standard PLP features are 'mapped' to the root UBM feature space by first classifying the channel based on eight channel-GMMs using a maximum likelihood criterion, and then shifting the features back by the exact amount by which the channel specific GMM mean was adapted away from the root UBM. This transformation is carried out on a per-frame basis, using the most occupied Gaussian mixture for determining the shift of the mean. Mapped features are then normalized to zero mean an unit variance per utterance. The mapped features are used as input for both the GMM and SVM subsystems.

The TNO 2005 GMM system uses as UBM the same 2048 mixture root UBM, which is no longer gender or channel specific. We used $r_m = 16$ in speaker MAP adaptation, adapting only means.

The TNO 2005 SVM subsystem uses a Generalized Linear Discriminant Sequence kernel [7]. The $N_f = 26$ features are expanded to a higher dimension by calculating all monomials up to order 3, resulting in $\sum_{n=1}^{3} \binom{N_f+n-1}{n} = 3653$ features per frame. A diagonal sums-of-squares matrix R was trained on all expanded background data, for which we used the same material on which the root UBM was trained. An SVM speaker model was trained by averaging all expanded features over time, and scaling the average by the inverse square root of the R matrix diagonal elements corresponding to the individual expanded feature dimension. These average expanded features were targeted the value '$+1$' in the SVM training procedure. All background speaker utterances were targeted '-1' in the SVM training tool SVMTorch from IDIAP [11].

2.2. Adaptation strategies

2.2.1. GMM adaptation

We used the same unsupervised, online adaptation strategy as described in [2], which is compliant with the NIST evaluation rules [12]. These state that the list of trials in the evaluation must be processed in the order given, and that the state of all (background) models of the system must be reset to the original state when a new model speaker is encountered in the evaluation. The trials are ordered grouping trials with the same target model, and ordering test segments within these groups by recording date. We use two parameters, an adaptation threshold a and an adaptation relevance factor r, to control the adaptation of models during an evaluation run. When the T-normalized [13] score exceeds a, we assume that the test segment is spoken by the target speaker, and we then further adapt the current speaker model using the test data by MAP adaptation using the relevance factor r. Proper values for a and r are determined from development testing.

We have also experimented with a technique in which we re-train the speaker models from scratch by adapting the UBM using all available speech material from the speaker, namely training speech and all observed test speech with a score exceeding the adaptation threshold. However, this technique appeared to perform less well than the first technique.

2.2.2. SVM Adaptation

A new approach presented here, is the adaptation of the SVM model to test speech data. Because we are not aware of a method of 'adapting' an SVM model using addition data, such as MAP adaptation in GMM's, we use the second approach described above for GMM. When the T-normalized score exceeds a threshold a, we train a new SVM based on all available training and test speech. This method has only one parameter a, and we have no parameter to control the 'aggressiveness' of the adaptation to new data.

2.3. Experimental results

2.3.1. Development for SRE 2004

Using the NIST SRE 2002 data and our 2004 GMM subsystem, we could obtain good adaptation results. In Figure 1 we have plotted the DET curves for male and female speakers separate, using both the reference and the adapting system. These results are obtained using $a = 3$ and $r = 6$ which optimize the $C_{\mathrm{DET}}^{\mathrm{min}}$, the mimimum Detection Cost [12]. The optimum threshold is the same as reported in [2]. It is interesting to note here, that the optimal relevance factor seems to be dependent on the relevance factor r_m used to obtain the original speaker models from the UBM. The optimum appears to be $r = 1.5r_m$. For our other 2004 GMM subsystems, which were trained with $r_m = 16$, we found an optimum $r = 24$. The drop in minimum DCF ($C_{\mathrm{DET}}^{\mathrm{min}}$) from 0.316 to 0.216 for the female speakers is quite spectacular, so we had high expectations of the SRE 2004 results.

2.3.2. Evaluation on SRE 2004

TNO submitted 4 adapted subsystems for the '1side-1side' condition, and 1 adapted subsystem to all other one-speaker conditions. Except for subsystem 2 1side-1side (shown in Figure 2), all other submissions did not show an improved performance over the unadapted condition. In Table 1 some of the results are summarized in terms of minimum and actual DCF, and equal error rate (EER).

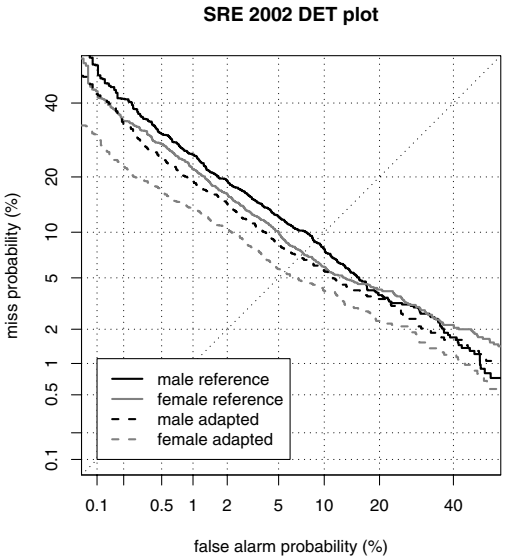

Figure 1: *results for the TNO 2004 subsystem 2 on the NIST SRE 2002 evaluation, separated for male and female speakers.*

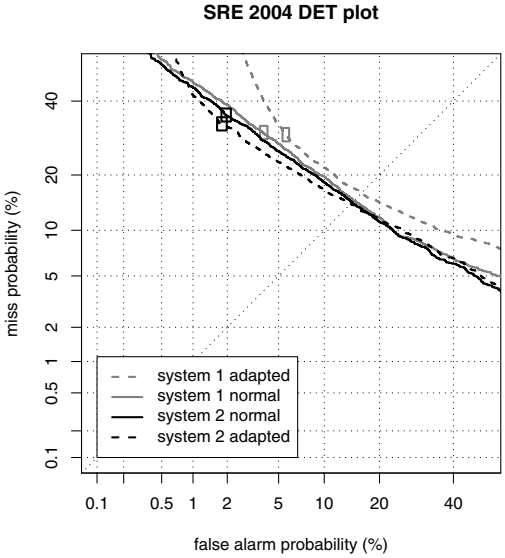

Figure 2: *SRE 2004 DET plots for selected TNO subsystems, for the '1side-1side' condition. The boxes indicate the actual decision operating points.*

2.3.3. Development for SRE 2005

For the NIST SRE 2005 we have the big advantage that there is an abundance of development test material that closely matches the evaluation conditions. For developing our SRE 2005 system, we have split the SRE 2004 material per gender in two separate groups of model speakers. The selection of speakers was made random, but under the condition that the EER and $C_{\mathrm{DET}}^{\mathrm{min}}$ points of both our subsystem 2 (GMM) and 5 (SVM) changed only slightly (0.01 and 0.1 %, respectively). The first split was used for T-norm model building, the second for development test performance measurements. We have carried out optimization of a and r for both the '1side-1side' and '8sides-1side' condition. For the GMM adaptation strategy, we found

Table 1: *Selection of results (all '1side' test condition) of the TNO submission to NIST SRE 2004, for the adapted and unadapted GMM systems. In italics are results for which adaptation improved performance.*

subsys	train	adapt	$C_{\mathrm{DET}}^{\mathrm{min}}$	C_{DET}	EER
1	10sec	-	0.880	0.914	27.8
1	10sec	•	0.987	1.047	*26.7*
1	30sec	-	0.673	0.738	19.4
1	30sec	•	0.919	0.955	21.0
1	1side	-	0.556	0.693	14.8
1	1side	•	0.784	0.858	16.1
2	1side	-	0.533	0.552	14.5
2	1side	•	*0.504*	*0.511*	*14.2*
3	1side	-	0.569	0.621	16.7
3	1side	•	0.657	0.704	*16.0*
4	1side	-	0.604	0.805	17.0
4	1side	•	0.944	1.110	19.9
1	3sides	-	0.494	0.737	13.0
1	3sides	•	0.766	0.924	14.8
1	8sides	-	0.455	0.851	13.0
1	8sides	•	0.762	0.938	13.7
1	16sides	-	0.367	0.990	11.1
1	16sides	•	0.737	0.950	12.8

an optimum at $(a, r) = (4, 16)$ for 8sides-1side evaluation condition, lowering $C_{\mathrm{DET}}^{\mathrm{min}}$ from 0.285 to 0.258. For the 1side-1side condition we could not find a parameter combination that improved the performance at the $C_{\mathrm{DET}}^{\mathrm{min}}$ operating point. We have no indication why this is the case, as the unadapted TNO 2005 GMM subsystem performs better ($C_{\mathrm{DET}}^{\mathrm{min}} = 0.450$) than the best 2004 subsystem ($C_{\mathrm{DET}}^{\mathrm{min}} = 0.523$ for the same subset or SRE 2004).

In the SVM adaptation approach we could not find a parameter setting which actually improved the $C_{\mathrm{DET}}^{\mathrm{min}}$, for the 1side-1side condition, either. For the 8sides-1side condition, however, we did find an optimum at $a = 4$, lowering $C_{\mathrm{DET}}^{\mathrm{min}}$ from 0.297 to 0.277. Adaptation results for both GMM and SVM systems are not as spectacular as for the SRE 2002 development test.

2.3.4. 'Oracle' adaptation

In order to investigate if adaptation on SRE 2004 material has any potential, we have made some 'oracle adaptation' runs [2]. This is an adaptation run by which the decision to adapt to a test segment is not based on the trial score, but according to the true speaker identity. This does improve GMM $C_{\mathrm{DET}}^{\mathrm{min}}$ considerably, down to 0.349 and 0.204 for 1side and 8sides training condition, respectively, with $r = 24$. Also with the SVM adaptation methodology $C_{\mathrm{DET}}^{\mathrm{min}}$ drops from 0.477 to 0.361 for 1side training, and down to 0.204 for 8sides training.

3. Discussion

It seems quite remarkable that automatic speaker adaptation does not work so well for SRE 2004 as it did for SRE 2002. The most important reason for the difference is most likely the different distribution of target speakers in the test. In Figure 3 we have plotted a histogram of the number of target trials per model for the NIST SRE of 2002 and 2004.

The histogram shows that for SRE 2004 there wasn't much opportunity to do adaptation, compared to SRE 2002. It is therefore more difficult to find the balance between 'false adaptation' (worsening the DET) and 'missed adaptation' (not improving). However, even if we only select speakers with more than 5 target trials in the evaluation, we don't see a strong benefit from

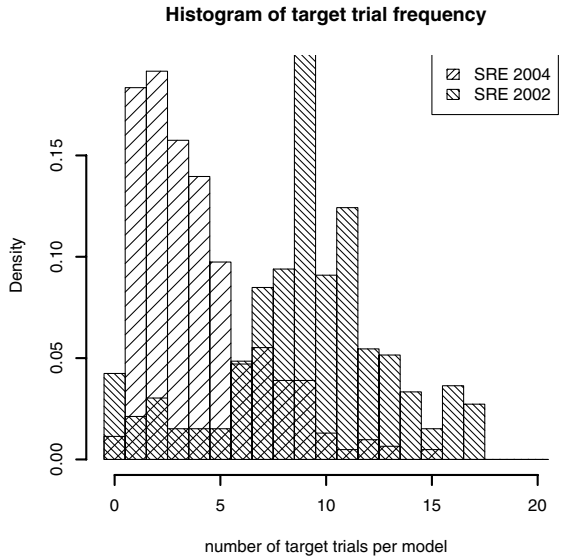

Figure 3: *Difference of the number of target trials per speaker model, between SRE 2002 and 2004.*

adaptation. The question remains why we saw some benefit from adaptation in the 1side-1side condition with our 2004 system, which we can't reproduce with a—much better baseline—2005 system. A possible explanation is that our 2005 system has much better channel adaptation, and that part of the effective adaptation seen in Figure 2 is due to channel or 'database collection' adaptation.

For the SRE 2002 evaluation priors the threshold a turned out to be close to the minimum DCF threshold, which is ~ 3, while for SRE 2004 it appears we have to be more conservative, $a = 4$. Notice that adaptation in general appears to raise the score threshold at which $C_{\mathrm{DET}}^{\mathrm{min}}$ operates, which needs more attention [14].

The difference in adaptation effect between 1side and 8sides training conditions suggest that there is a minimum level of performance necessary to obtain any improvement in such target-scarce conditions. In that sense it didn't help that the 2004 edition was 'harder' than previous years, due to a variety of reasons. The 2004 data was obtained from previously unseen data collection, and all trials involve 'different number' trials, punishing sensitivity to channel effects. Further, many trials involve different spoken languages between training and testing, which was new in 2004.

The main reason that most of the adaptation runs in SRE 2004 worsened performance, is the bad calibration (choice of a decision threshold) for many conditions, as can be seen from the difference between actual detection cost C_{DET} and $C_{\mathrm{DET}}^{\mathrm{min}}$. The reason for this is probably that 'unmatched T-norm' models were used with fixed decision thresholds. We used T-norm models trained on approximately 2-minute conversation excerpts. Longer model training seems to narrow the imposter score distribution, which leads to higher T-normalized scores. Working with a fixed adaptation threshold $a = 3$ is therefore penalized for all badly calibrated conditions.

For well enough baseline performance we were able to see some performance improvement using adaptation for the SVM system. The adaptation in this condition is not very aggressive, because in our implementation the 8sides training lead to 8 positive targets for the SVM, while each accepted test utterance adds only one target. With only few targets, this doesn't change the models dramatically.

We have fused the 2005 systems GMM and SVM results, resulting to an adapted $C_{\mathrm{DET}}^{\mathrm{min}} = 0.231$ compared to a fused reference value of 0.259. This fusing is performed using separate decisions for SVM and GMM adaptation; we have not tried to base the adaptation decision on a fused result.

The NIST SRE 2005 will show if our adaptation system is better calibrated this year. The effectiveness of test segment adaptation depends strongly on the evaluation target priors, but it is of course essential to proper evaluation that these are unknown at the time of evaluation.

4. References

[1] N. Mirghafori and L. Heck, "An adaptive speaker verification system with speaker dependent a proiri deciasion thresholds," in *Proc. ICSLP*, 2002, pp. 589–592.

[2] C. Barras, S. Meigner, and J. L. Gauvain, "Unsupervised online adaptation for a speaker verification system over the telephone," in *Proc. Speaker Odyssey*, 2004.

[3] M. Przybocki and A. Martin, "NIST speaker recognition evaluation chronicles," in *Proc. Odyssey 2004 Speaker and Language recognition workshop*. ISCA, June 2004, pp. 15–22.

[4] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech 1997*, Rhodes, Greece, 1997, pp. 1895–1898.

[5] N. Brümmer, "Application-independent evaluation of speaker detection," in *Proc. Odyssey 2004 Speaker and Language recognition workshop*. ISCA, June 2004, pp. 33–40.

[6] M. Hebert and N. Mirghafori, "Desperately seeking impostors: Data-mining for competitive impostor testing in a text-dependent speaker verification system," in *Proc. ICASSP*, 2004, pp. 361–364.

[7] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. ICASSP*, 2002, pp. 161–164.

[8] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[9] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*. Crete, Greece, 2001.

[10] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP*, 2003, pp. 53–56.

[11] R. Collobert, S. Bengio, and J. Mariéthoz, "Torch: a modular machine learning software library," IDIAP, Tech. Rep. IDIAP-RR 02-46, 2002.

[12] "The NIST year 2004 Speaker Recognition Evaluation Plan," http://www.nist.gov/speech/tests/spk/2005/index.htm, 2004.

[13] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independetn speaker verification systems," *Digital Signal Processing*, 2000.

[14] N. Mirghafori and M. Hébert, "Parameterization of the score threshold for a text-dependent adaptive speaker verification system," in *Proc. ICASSP*, 2004, pp. 365–368.