Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006

Niko Brümmer, Lukáš Burget, *Member, IEEE,* Jan "Honza" Černocký *Member, IEEE,* Ondřej Glembek, *Student Member, IEEE,* František Grézl, *Member, IEEE,* Martin Karafiát, *Member, IEEE,* David A. van Leeuwen, Pavel Matějka, *Member, IEEE,* Petr Schwarz, *Member, IEEE,* Albert Strasheim

***Abstract*—This paper describes and discusses the 'STBU' speaker recognition system, which performed well in the NIST Speaker Recognition Evaluation 2006 (SRE). STBU is a consortium of 4 partners: Spescom DataVoice (South Africa), TNO (The Netherlands), BUT (Czech Republic) and University of Stellenbosch (South Africa). The STBU system was a combination of three main kinds of sub-systems: (1) GMM, with short-time MFCC or PLP features, (2) GMM-SVM, using GMM mean supervectors as input to an SVM, and (3) MLLR-SVM, using MLLR speaker adaptation coefficients derived from an English LVCSR system. All sub-systems made use of supervector subspace channel compensation methods—either eigenchannel adaptation or nuisance attribute projection. We document the design and performance of all sub-systems, as well as their fusion and calibration via logistic regression. Finally, we also present a cross-site fusion that was done with several additional systems from other NIST SRE-2006 participants.**

***Index Terms*—Speaker recognition, GMM, SVM, eigenchannel, NAP, Fusion.**

I. Introduction

This paper documents significant elements of the state-of-the-art in text-independent telephone speaker recognition, as measured in the NIST Speaker Recognition Evaluation 2006 (SRE), via a description of the design and performance of the 'STBU' submission. It expands on a short paper published at ICASSP [1]. The U.S. National Institute of Standards and Technology (NIST) organizes yearly SRE evaluations [2], [3] to contribute to the direction of research efforts and to calibrate the technical capabilities of different academic and industrial sites active in text-independent speaker recognition.

The STBU submission to the NIST SRE-2006 was the result of a collaboration between four institutes:

- Spescom DataVoice (SDV), South Africa,
- TNO, The Netherlands,
- Brno University of Technology (BUT), Czech Republic, and
- University of Stellenbosch (SUN), South Africa.

The STBU consortium was formed to learn and share the technologies and available know-how among partners. Another, equally important, reason to join efforts was that most successful submissions to NIST evaluations fuse the results of several sub-systems to decrease error rates. Simply put, a consortium can generate more diverse systems, and even if the theoretical base is very similar, subtle details in implementation, features, background models, channel normalization and training can make the fused system more accurate.

Admittedly, this paper is not for novices in speaker recognition. Rather, it assumes familiarity with basic approaches such as Universal Background Model-Gaussian Mixture Modelling (UBM-GMM) [4], sequence kernel Support Vector Machines [5] and more advanced channel compensation approaches such as Eigenchannel Adaptation [6] and Nuisance Attribute Projection (NAP) [7]. Further, the reader is assumed to be familiar with the NIST SRE-2006 task of speaker detection [8] and specifically with the '1conv4w-1conv4w' condition[1], where a *detection trial* consists of a pair of speech segments, and where the objective of the exercise is to decide independently for each of several thousand trials, whether the two segments were spoken by the same speaker, or by two different speakers. *Speech segment* here denotes an excerpt of approximately 5 minutes, from one of the 2 channels of a 4-wire recording of a telephone conversation between two people.

The paper is organized as follows: Section II presents the basic system types grouped into three categories. Section III presents the systems from different STBU sites in more detail. In Section IV, we describe in detail the theory and implementation of system fusion and calibration using logistic regression. In particular, we discuss how calibration was done to meet both the traditional C_{det} and the new C_{llr} metrics. Results are presented in Section V—this section also analyzes language dependence which was an important issue in SRE-2006. Finally, Section VI presents a cross-site fusion of STBU

The authors appear in alphabetical order

Niko is with Spescom DataVoice, Stellenbosch, South Africa and with University of Stellenbosch.

Pavel, Lukáš, Petr, Ondřej, Martin, František and Honza are with Speech@FIT, Faculty of Information Technology Brno University of Technology, Czech Republic.

David is with TNO Human Factors, Postbus 23, 3769 ZG Soesterberg, The Netherlands.

Albert is with University of Stellenbosch, Department of Electrical and Electronic Engineering, Stellenbosch, South Africa.

[1]For details see the evaluation plan, via http://www.nist.gov/speech/tests/spk/2006/

sub-systems together with several systems from other SRE-2006 participants. We conclude the paper in Section VII.

II. System description

We used three basic system types: Eigenchannel GMM, GMM-SVM and MLLR-SVM. All sub-systems had in common that they used one of two forms of linear supervector subspace channel compensation technique: (i) For *eigenchannel adaptation*, supervectors were extracted from GMM models, compensated for channel effects, translated back to adapted GMM models and then employed in the usual way to score the tests. (ii) In the case of the SVM-based systems, supervectors were extracted either from GMMs or from MLLR adaptation coefficients and were then subjected to *nuisance attribute projection* to cancel channel effects. Following that, the supervectors are employed in the usual way to train SVM models which can be scored against test supervectors. More detail follows below.

A. Common signal processing

All sub-systems used standard features such as Mel-frequency cepstral coefficients (MFCC) or perceptual linear prediction (PLP) features. The basic cepstral features were augmented with derivatives up to third order. A set of several frame selection criteria were applied: (a) frame energy must be more than than 30 dB below the maximum frame energy; (b) frame energy at least 3 dB above energy in other channel (cross channel squelch); (c) segmentation from BUT's Hungarian phone recognizer; (d) strongly voiced syllable nuclei detector; (e) ASR word transcript segmentation provided by NIST. RASTA (relative spectral) filtering [9], short-time Gaussianization [10] and heteroscedastic linear discriminant transformation (HLDA) [11], [12] were used for basic channel normalization, feature decorrelation and dimensionality reduction.

B. Feature mapping

TNO and BUT used the channel-compensation technique of feature mapping [13] to post-process all of their acoustic features. However, post-evaluation experiments by BUT [14] strongly suggest that when eigenchannel or NAP channel compensation are used, then feature mapping becomes unnecessary.

In the BUT systems, 8 feature mapping channels were found by unsupervised iterative re-clustering of conversations [15], primed with the TNO feature mapping labels (CDMA, GSM, carbon button, electret per gender), as used in SRE-2005. These were augmented with 6 channels determined from SRE-2004 labels (cellular, cordless, standard per gender). The TNO feature mapping used 16 classes, and was trained with balanced quantities from Switchboard (640 speakers) and Fisher (1000 speakers) databases.

C. Eigenchannel GMM

We adopted the term 'eigenchannel' as used in speaker recognition from Kenny [6]. It was introduced to the NIST SRE by SDV in 2004 [16], revisited by Kenny [17], [18] and Vogt [19] in SRE 2005, and again by several sites in various forms in SRE-2006 [20].

In our Gaussian mixture model (GMM) system [14], speaker models were trained in the usual way by adapting from a universal background model (UBM [4]) by maximum a-posteriori (MAP) adaptation [21]. Only means of Gaussian components are adapted.

In the following, we will use the notion of *supervectors*[2]: Since our GMMs differ only in means, each model can be represented by the concatenation of all the mean vectors of all the Gaussians in the model. (We normalized each mean by the corresponding standard deviation.)

In eigenchannel adaptation, a model that has been trained under one channel condition, may be adapted towards a different channel condition of new test data, to reduce mismatch when the speaker is the same. Importantly, the adaptation must be constrained so that adaptation between different speakers is suppressed. This constraint is effected by adapting GMM models in supervector space, but only in a very small[3] subspace.

The adaptation is effected by maximizing (with a single iteration of the Expectation Maximization (EM) algorithm [21]) the MAP-criterion, $P(\{f_t\}|\mathbf{m} + \mathbf{V}\mathbf{x})P(\mathbf{x})$, w.r.t. the low-dimensional 'channel mismatch' vector $\mathbf{x}$ [16], [14]. Here, $\{f_t\}$ is the sequence of acoustic feature vectors in the test segment, $\mathbf{m}$ is the supervector representing the original model, $\mathbf{V}$ is a low-rank matrix that spans the adaptation subspace, and $P(\mathbf{x})$ is a zero-mean, unit-covariance Gaussian prior on the channel mismatch. In later experiments, we found the prior to be unimportant and that the MAP-criterion could be replaced by a simpler ML-criterion, by ignoring the prior. The adaptation subspace $\mathbf{V}$ was trained via the same eigen-analysis that was used to find the NAP-subspace, see Section II-F1.

In the variant of this system without T-norm (test normalization), the score for each trial was calculated as $\log P(\{f_t\}|\mathbf{m}_a) - \log P(\{f_t\}|\mathbf{U}_a)$, where $\mathbf{m}_a$ and $\mathbf{U}_a$ are the independently adapted target and universal background models. In the T-normed variant, the score was normalized in the usual way [22], but with each T-norm model also independently adapted. The EM-algorithm for adaptation of multiple T-norm models was streamlined by using the state occupancy probabilities of the UBM for all models, as proposed by [19].

The BUT eigenchannel GMM system and its interaction with various feature-space compensations such HLDA and feature mapping is analysed in more detail in [14].

D. GMM-SVM

In this type of system, GMM supervectors, as described in the previous section, are extracted not only from target-model training speech segments, but also for all other background and test speech segments. In other words, each speech segment (conversation side) is represented by a single GMM

[2]Supervectors are just rather large vectors, where 'super' serves to distinguish them from the much smaller short-time feature vectors.

[3]In this case the subspace was 30-dimensional while the full supervector dimension was almost 80000.

supervector. The target and background supervectors are then used to train support vector machine (SVM) speaker models against which the test supervectors are scored [23], [24]. The SVM uses a linear kernel in supervector space. Each SVM is trained using the single available positive example from the target speaker, and many[4] negative examples from a pool of background speakers.

All our SVM sub-systems used NAP as a preprocessing step before SVM training. This is described in Section II-F.

E. MLLR-SVM

This type of system makes use of large vocabulary continuous speech recognition (LVCSR). Previous work [25] has already shown that the adaptation matrices that LVCSR systems use to adapt towards new speakers are excellent features for speaker recognition.

The sub-systems in this paper used the coefficients from constrained maximum likelihood linear regression (CMLLR) and maximum likelihood linear regression (MLLR) transforms, as estimated by the LVCSR system developed in AMI project[5] submitted to NIST Rich Transcription 2005 evaluations [26]. This adaptation was 'supervised' by using the ASR transcripts[6], as made available by NIST for all speech data in SRE-2005 and 2006. Since NIST did not provide pronunciation dictionary, we used the AMI dictionary and we generated the missing pronunciations automatically. With this, we were able to generate the triphone alignment, to apply vocal tract length normalization (VTLN) and to estimate the coefficients of CMLLR and MLLR transformations.

These coefficients were normalized and concatenated into supervectors and then used with SVMs, exactly as described in the previous subsection for the GMM supervectors.

F. Nuisance attribute projection (NAP)

All of our SVM sub-systems used NAP [7], [27] to remove unwanted channel or inter-session variability. There are different ways in which the NAP transform may be estimated and applied. We give here the general recipe that we applied in all of the STBU SVM systems. We also show how the eigenchannel adaptation matrix $\mathbf{V}$ was obtained.

1) NAP training: The data collection used in SRE-2004 was specifically designed to contain a large channel variability. Hence, as training material for the NAP-transforms we used whole conversation sides from the NIST SRE-2004. This data includes circa 310 speakers for most of which there are 10 or more conversation sides, or *sessions*. The steps for estimating the NAP transform are:

- Extract a supervector of dimension[7] $D_{\rm sv}$ for each session of each speaker.
- For each speaker, calculate the mean supervector over all of the available supervectors of that speaker. Then subtract the mean from all of the vectors for that speaker. Pooling all these difference vectors then gives a large matrix $\mathbf{D}$ of supervectors from which most of the speaker variability has been removed, but where the inter-session (or nuisance) variability remains. The matrix $\mathbf{D}$ has dimensions $D_{\rm sv} \times N_{\rm ses}$, where $N_{\rm ses}$ is the total number of sessions.
- Select the NAP transform dimension, denoted as $D_{\rm NAP}$. We typically used $D_{\rm NAP} = 40$, but this dimension should be chosen empirically as the one which gives best results.
- Now perform a principal component analysis (PCA) on $\mathbf{D}$. That is, we need to find the $D_{\rm NAP}$ principal eigenvectors of the normalized scatter matrix[8] $\frac{1}{N_{\rm ses}}\mathbf{D}\mathbf{D}^T$. Since the number of session vectors is typically several thousand, and the supervector dimension can be in the tens of thousands, some careful engineering may be needed to find these eigenvectors on machines of limited memory and CPU capacity. Some hints are given in Section II-H. We denote the $D_{\rm sv} \times D_{\rm NAP}$ matrix of principal eigenvectors as $\mathbf{E}$.
- Since an iterative eigenvector algorithm typically gives approximate solutions, it is a good precaution to normalize and mutually orthogonalize the columns of matrix $\mathbf{E}$, for example by *singular value decomposition* (SVD) of $\mathbf{E}$. If the eigenvectors are not orthonormal, the NAP-transform fails to project the nuisance subspace away completely.

2) Eigenchannel matrix: If the ML-version (without channel mismatch prior) of eigenchannel adaptation is used, it suffices to simply set $\mathbf{V} = \mathbf{E}$, where $\mathbf{V}$ is the matrix mentioned in Section II-C. However if MAP-adaptation is used, then each column j of $\mathbf{V}$ should be scaled by $\sqrt{2e_j}$, where e_j is the corresponding eigenvalue. (Directions in nuisance subspace with relatively smaller variances are thereby allowed to adapt to a lesser extent.)

3) NAP-projection: Once the orthonormal[9] NAP-subspace $\mathbf{E}$ has been trained as explained above, we may use it to train SVM speaker models that are more robust against inter-session variability. The basic NAP-transform is designed to be applied with linear-kernel SVMs. The transform must be applied to all supervectors (target and background) before they are used in SVM model training. That is, each supervector $\mathbf{v}$ is transformed as:

$$\mathbf{v}' = \mathbf{v} - \mathbf{E}(\mathbf{E}^T\mathbf{v}), \tag{1}$$

where ^{T} denotes transpose. By orthonormality, this transformation is idempotent [27]. This means it is not necessary to also NAP-transform the test supervectors[10], before they are scored against the SVM models. Finally, note that the NAP transform should be applied *before* SVM training. It does not

[4]Background size was of the order of 2000, which is much smaller than the supervector dimension. In practice this always results in SVM models with zero training errors. This makes selection of the SVM regularization constant irrelevant.

[5]See http://www.amiproject.org

[6]from a different English LVCSR system

[7]For GMM supervectors the dimension is the acoustic feature dimension times the number of GMM components. Numerical values are given in Table I.

[8]$\mathbf{D}$ has zero mean, so that this normalized scatter matrix acts as estimate of within-speaker covariance.

[9]$\mathbf{E}^T\mathbf{E} = \mathbf{I}$

[10]It would also not matter if this operation *was* repeated because of the idempotence.

help to apply the NAP-transform afterwards to test vectors or to models that have been trained on unprojected data.

G. Division of training data

Although not all STBU sites had the same speech databases at their disposal, a general division of training data was made early in the design stage to which all sites adhered. Starting with the most recent collection, we used: *SRE-2005* exclusively for sub-system development testing, calibration and fusion; *SRE-2004* for eigenchannel, NAP, UBM, T-norm and rank normalization; *SRE 1999–2003, Fisher, Switchboard* for UBM training, feature mapping, SVM background, and T-norm.

H. Some notes on computational efficiency

For experiments with these complex systems and large test databases it is important to have fairly efficient implementations of the various algorithms. Here we give some hints:

- Store the top-N Gaussian index for each speech frame, where typically $N = 5$ [4]. Note that for obtaining this index for a frame f_t, only the distance to the Gaussian centers needs to be evaluated, and the exponentiation can be postponed or even omitted in the GMM-SVM case.
- For MAP adaptation of GMM means, only the top-N Gaussian components need to be evaluated in the 'expectation-step,' making this typically a factor 100 faster [16]. Since this needs to be performed for each test segment (in the GMM-SVM case), this makes a big difference.
- In the estimation of the NAP projection, rather than calculating the principal D_{NAP} eigenvectors of $\mathbf{D}\mathbf{D}^T$, calculate the principal eigenvectors of $\mathbf{D}^T\mathbf{D}$ (which is much smaller), and left-multiply these by $\mathbf{D}$ afterwards.
- Using ARPACK or Matlab's `eigs()`, explicit calculation of $\mathbf{D}^T\mathbf{D}$ is not necessary, but rather a function $f(\mathbf{x}) = \mathbf{D}^T\mathbf{D}\mathbf{x}$ can be provided. This function can be calculated without transposing large matrices using $f(\mathbf{x}) = \left((\mathbf{D}\mathbf{x})^T\mathbf{D}\right)^T$.
- For training SVM models (e.g., using `libSVM` [28]), pre-compute the Gram (kernel) matrix between all background speakers. Then for each new target/T-norm speaker, only one row and column needs to be replaced in the Gram matrix. This speeds up SVM training with orders of magnitude.
- For SVM scoring, SVM models can be folded, or compacted [5], into a single vector by calculating a weighted sum of the support vectors. Evaluation of a score is then just calculation of an inner product and T-normalization is just a matrix-vector multiplication.

III. Sub-systems and their diversity

In the fusion of sub-systems, we found it advantageous to include in each fusion several very similar, but not identical, systems. Indeed, in post-evaluation experiments we found that leaving any of the sub-systems out caused significant deterioration in performance. These sub-systems were different because each was built by a different team, using different front-ends, different development databases and somewhat different flavours of the subspace channel compensation techniques. See Table I for a summary of the main characteristics of the various sub-systems.

Some remarks not captured in the table are the following. In an attempt to compensate for asymmetric system design, SDV provided two similar sub-systems: A *reverse* system swapped test and train speech segments for each trial, but was otherwise the same as the *forward* system. Because one speech segment is used for training the model and the other for obtaining a score, this swapping makes the system more symmetric. Experiments have shown that fusing these to sub-systems leads to better performance.

The acoustical features from BUT, as well as the MLLR transform data, were used by SUN. SUN provided two versions of the MLLR-SVM system, differing in the number of MLLR transforms.

CMLLR and MLLR transforms were trained for each speaker. At first, CMLLR was trained with two classes (speech + silence). On top of it, MLLR with two (SUN) or three (BUT, SUN) classes (the two speech classes were obtained by automatic clustering on the LVCSR training data + silence) was estimated. Using more classes caused missing data problems for some files, and was found not to lead to better performance. Both CMLLR and MLLR transform matrices were estimated as block-diagonal in 13-coefficient wide streams.

IV. Fusion, calibration and decisions

The crux of the STBU design was to *fuse* multiple sub-systems into a single effective system. By fusion we mean the following: Let x represent a speaker detection trial[11] and let this trial be processed in parallel by N sub-systems, each of which produces a real-valued output *score*, where more positive scores favour the target hypothesis (same-speaker) and more negative the non-target hypothesis (different-speakers) . The score of the ith sub-system is denoted $s_i(x)$. These scores are fused using linear combination:

$$s_{\mathrm{f}} = s(x, \mathbf{w}) = w_0 + \sum_{i=1}^{N} w_i s_i(x) \tag{2}$$

where s_{f} is the fused output score and $\mathbf{w} = [w_0, w_1, \ldots, w_N]$ is a vector of real-valued weights. Perhaps counter intuitively, some of the weights may be negative.

A. Logistic regression

The fusion weights were obtained by *logistic regression* [29] training on a database of supervised scores. We used all 1conv4w-1conv4w trials of the NIST SRE-2005 for this purpose. It is important to note that all development of the sub-systems did not make use of any 2005 data. If for example, 2005 data had been used to train NAP/eigenchannel, then the scores produced by these systems on the same data would have been over-optimistic and therefore not suitable for training fusion and calibration weights.

[11] Recall a trial consists of two speech segments.

TABLE I
SUMMARY OF ALL SUB-SYSTEMS COMPONENTS. LEGEND TO DATA SOURCE: SW: SWITCHBOARD, Snn: NIST SRE-'nn, F1: FISHER RELEASE 1. FRAME SELECTION METHODS (A)–(E) ARE EXPLAINED IN SECTION II-A.

Site	SDV	BUT			SUN		TNO
System	GMM-SVM	GMM	GMM-SVM	MLLR-SVM	GMM-SVM	MLLR-SVM	GMM-SVM
Features	12 MFCC, Δ	12 MFCC+C_0, Δ^3		12PLP+C_0,Δ^3	12 MFCC+C_0, Δ^3	12PLP+C_0,Δ^3	12 PLP + $\log E$, Δ
HDLA dimension		39	39	39	39	39	
Frame selection	(b),(d)	(a)–(c)	(a)–(c)	(e)	(a)–(c)	(e)	(a)
N_f	24	39	39	39	39	39	26
UBM sources	S99–S03	S04	S04		S04		SW, S01–S03, F1
N_G	512	2048	512		512		512
Feature mapping channels		14	14		14		16
Relevance factor	16	19	19		19		16
D_{sv}	12288	79872	19968	1638	19968	1092, 1638	13312
D_{NAP}	40	30	40	15	40	15	40
SVM Background speakers	> 2000		2866	310	2606	310	1640
source	S99–S03		F1, S02	S04	F1	S04	SW, S01–S03, F1
T/Rank norm speakers	T: 310	T:260	T: 1080		T: 300	T: 310	T: 310
			R: 2866	R:310			
source	S04	S02	F1, S02	S04	F1	S04	S04

The aim of logistic regression training is two-fold: First, it should improve *discriminative* ability, i.e., the DET-curve of the fused system should be better than the DET-curves of all the input systems. This is clearly demonstrated in Figs. 2, 3 and 6, which compare DET-plots of sub-systems against their fusion. Secondly it should *calibrate* the output score, so that it functions as a well-calibrated *log-likelihood-ratio*. That is, the training strives to achieve

$$s_{\text{f}} \approx \log \frac{P(s_{\text{f}}|H_{\text{tar}})}{P(s_{\text{f}}|H_{\text{non}})} \tag{3}$$

where H_{tar} and H_{non} denote target and non-target hypotheses respectively [30]. With a linear fusion such as (2), the degrees of freedom, which may be adjusted to optimize calibration, effectively form an affine transform—i.e., scores can be scaled and shifted. Scaling and shifting of scores does not affect discrimination and does not change the DET-plot.

There is a subtle difference between our use of logistic regression and the way in which it is traditionally applied in many other pattern recognition problems [31]. As mentioned, we train the fused score to function as a *log-likelihood-ratio*, while in other problems it is appropriate to train the score to function as *posterior log-odds*:

$$s'_{\text{f}} \approx \log \frac{P(H_{\text{tar}}|s'_{\text{f}})}{P(H_{\text{non}}|s'_{\text{f}})} = \log \frac{P_{\text{tar}}}{1-P_{\text{tar}}} + \log \frac{P(s'_{\text{f}}|H_{\text{tar}})}{P(s'_{\text{f}}|H_{\text{non}})} \tag{4}$$

In other words, the traditional posterior log-odds, s'_{f} and our log-likelihood-ratio, s_{f}, differ essentially in an additive term, namely the *prior log-odds*,

$$\text{logit}\, P_{\text{tar}} = \log \frac{P_{\text{tar}}}{1-P_{\text{tar}}} \tag{5}$$

As is shown below, this is easily handled by a small modification of the traditional logistic regression objective function. Let $\mathcal{X}_{\text{tar}}$ and $\mathcal{X}_{\text{non}}$ respectively represent sets of target and non-target trials. Our logistic regression objective function is:

$$\begin{aligned}\mathcal{O}(\mathbf{w}, P_{\text{tar}}) = &\frac{P_{\text{tar}}}{\|\mathcal{X}_{\text{tar}}\|} \sum_{x \in \mathcal{X}_{\text{tar}}} \log(1 + e^{-s(x,\mathbf{w}) - \text{logit}\, P_{\text{tar}}}) \\ &+ \frac{1-P_{\text{tar}}}{\|\mathcal{X}_{\text{non}}\|} \sum_{x \in \mathcal{X}_{\text{non}}} \log(1 + e^{s(x,\mathbf{w}) + \text{logit}\, P_{\text{tar}}})\end{aligned} \tag{6}$$

where $\|\mathcal{X}\|$ denotes the number of trials in set $\mathcal{X}$. Note that the objective is parameterized by the target prior P_{tar}. This adaptation of the logistic regression objective function allows one to set the parameter P_{tar} independently of the proportion of target trials in the training database, to match the target prior of an envisaged application of the fusion. Since the purpose of this fusion was to optimize for the NIST SRE C_{det} objective, we set [32]

$$\text{logit}\, P_{\text{tar}} = \text{logit}\, P'_{\text{tar}} + \log \frac{C_{\text{miss}}}{C_{\text{fa}}} \tag{7}$$

where $(P'_{\text{tar}}, C_{\text{miss}}, C_{\text{fa}}) = (0.01, 10, 1)$ are the parameters specified by the evaluation plan[12]. This gives $P_{\text{tar}} = 0.0917$. In experiments over a few different NIST SRE evaluation sets, we have found that, although performance of the logistic regression is relatively insensitive to the parameter P_{tar}, it does help to set it to the above value.

On the other hand, if the fusion is to be designed to optimize for the new C_{llr} objective [32], [33], which was adopted as a secondary evaluation objective in the most recent NIST SRE Evaluation plan[13], then it would be better to choose $P_{\text{tar}} = 0.5$. Indeed, if (6) is reformulated as a function of the scores, rather than of $\mathbf{w}$, then at $P_{\text{tar}} = 0.5$, it is just the C_{llr} objective.

At a fixed value of P_{tar}, the objective $\mathcal{O}(\mathbf{w}, P_{\text{tar}})$ is a convex function of $\mathbf{w}$, and it has a global minimum. This means it can be efficiently optimized with, for example, conjugate-gradient methods. We implemented a conjugate-gradient algorithm in Matlab, based on the work of Minka[14], but adapted to our variant of the objective function. This code is freely available as part of the FoCal toolkit[15].

B. Missing trials

We had the complication that not all sub-systems were able to contribute a score for each trial, because of failure to detect speech in training or test segment, or lack of transcription. This necessitated a two step fusion strategy:

[12]See http://www.nist.gov/speech/tests/spk/
[13]See http://www.nist.gov/speech/tests/spk/2006/.
[14]See http://www.stat.cmu.edu/~minka/papers/logreg/
[15]See http://www.dsp.sun.ac.za/~nbrummer/focal/

1) First, each sub-system on its own was subjected to an affine calibration transformation[16], also trained via logistic regression, with $P_{\rm tar} = 0.5$. This calibration gave the scores a log-likelihood-ratio interpretation. The training data for this calibration were all trials that each sub-system could contribute out of the SRE-2005 (1conv4w-1conv4w) trials.
2) Next, scores (log-likelihood-ratios) of *zero* were inserted for all missing trials. Now, all sub-systems had valid scores for all trials and the fusion could be trained as explained above.

C. Decisions

The beauty of a score that is calibrated so that approximation (3) holds is, that decisions with near-optimal expected cost can be made by using standard, theoretically determined score thresholds.

In past years, it was standard practice for NIST SRE participants to empirically determine score thresholds by optimizing average $C_{\rm det}$ performance over a database of supervised scores. This strategy indeed often worked well for the particular operating point defined via the $C_{\rm det}$ parameters. But if decisions at different operating points (different prior or costs) were required for applications other than the NIST SRE, then the threshold optimization procedure would have to be repeated.

The advantage of calibrated scores is that the empirical optimization, e.g. via logistic regression, over the supervised database needs to be performed once only. Thereafter, theoretical thresholds can be used to give good performance over a wide range of operating points [33]. If the goal is to make decisions that optimize $C_{\rm det}$, then the theoretical threshold is just the negative of (7):

$$\theta_{\rm DET} = -\operatorname{logit} P'_{\rm tar} - \log \frac{C_{\rm miss}}{C_{\rm fa}} \tag{8}$$

For $(P'_{\rm tar}, C_{\rm miss}, C_{\rm fa}) = (0.01, 10, 1)$, this gives $\theta_{\rm DET} = 2.29$. The decision rule is then:

$$\begin{aligned} s_{\rm f} \geq \theta_{\rm DET} &\mapsto \mathsf{accept}, \\ s_{\rm f} \leq \theta_{\rm DET} &\mapsto \mathsf{reject}. \end{aligned} \tag{9}$$

This new calibration-based strategy has indeed worked well, as demonstrated by small $C_{\rm det} - C_{\rm det}^{\rm min}$ discrepancies in the system submitted by SDV in the NIST SRE-2005, as well as for 5 of the best-performing systems[17] in the NIST SRE-2006, all of which used logistic regression-based calibration with a 2.29 threshold.

D. Non-linear calibration (STBU-3)

As mentioned above, we are concerned with optimizing the discriminative ability (DET-curves), as well as the calibration, or *actual decision-making ability* of our scores. The traditional evaluation tools which are applied to analyse NIST SRE results include both (i) DET-curves to analyse discriminative ability over a *wide operating range* and (ii) $C_{\rm det}$ to analyse actual decision-making ability at a *fixed operating point*. The new $C_{\rm llr}$ metric serves to fill this gap: It evaluates average actual decision-making ability of log-likelihood-ratio scores, over a *wide operating range*. For a tutorial introduction to $C_{\rm llr}$ see [32] and for a reference implementation to calculate $C_{\rm llr}$ and $C_{\rm llr}^{\rm min}$, see the above-mentioned FoCal toolkit.

With our submissions STBU-1 and STBU-3, we tried to optimize calibration performance respectively for the traditional $C_{\rm det}$ and for the new $C_{\rm llr}$. STBU-1 was a straight-forward linear fusion (2) optimized with logistic regression with the parameter $P_{\rm tar} = 0.0917$. As explained, this fusion effects an affine calibration transformation.

STBU-3 took the score the output, $s_{\rm f}$, of STBU-1 and then subjected it to a further non-linear calibration stage. That is, the score of STBU-3 was obtained by:

$$s_{\rm c}(s_{\rm f}) = \log \frac{\alpha(e^{s_{\rm f}} - 1) + 1}{\beta(e^{s_{\rm f}} - 1) + 1} \tag{10}$$

where $0 < \beta < \alpha < 1$. This is a strictly increasing sigmoid function, which saturates below at approximately $-\operatorname{logit} \alpha$ and above at approximately $-\operatorname{logit} \beta$. The parameters α and β are likewise found by optimizing the logistic regression objective, but here our aim was to optimize for $C_{\rm llr}$ rather than $C_{\rm det}$, so we set the parameter $P_{\rm tar} = 0.5$. Code for performing this optimization[18] is also available in the FoCal toolkit, as well as a derivation for the particular form of this saturating non-linearity.[19] As shown in Table IV in the results section, the STBU-3 strategy did indeed improve calibration as measured by $C_{\rm llr}$.

V. Results and discussion

A. Comments on individual systems

In the development of individual systems, many configurations and parameters were tested and it is not possible to cover everything in this paper. We will therefore concentrate on the most important findings. The results will be presented on DET plots on 2006 data in Fig. 1:

1) Compare the influence that eigenchannel adaptation has on the GMM system (left) to the influence of NAP on the GMM-SVM (middle), as both techniques have similar underlying principles. We have found that, while eigenchannel greatly helps in the GMM system (and actually makes feature mapping unnecessary [14]), NAP helps in the GMM-SVM but to a much smaller extent. We attribute this to the fact that linear-kernel SVM models orient the score projection axis approximately perpendicular to the subspace spanned by all the background supervectors, which also includes much channel variation.
2) NAP in the MLLR-SVM sub-system (right) also helps, but it seems that SVM itself is able to exploit the

[16]This is the same as a fusion with a single input.

[17]NIST SRE rules prohibit publishing explicit performance details of other participants.

[18]Because of the saturation, the objective function may become non-convex. This makes it harder to optimize and it may fail to converge if not appropriately initialized.

[19]See http://www.dsp.sun.ac.za/~nbrummer/focal/cllr/calibration/s_cal/derivation.pdf

speaker-discriminative information in LVCSR adaptation matrices to some extent.

B. Fused systems and their results

Three fused systems were submitted to the evaluation. The primary submission, STBU-1U (unsupervised adaptation mode) is an 11-fold fusion of:

1) GMM-SVM forward, T-normed (SDV)
2) GMM-SVM reverse, T-normed (SDV)
3) Eigen-channel GMM (BUT)
4) Eigen-channel GMM T-normed (BUT)
5) GMM-SVM T-normed (BUT)
6) MLLR3-SVM (BUT)
7) GMM-SVM T-normed (SUN)
8) MLLR2-SVM (SUN)
9) MLLR3-SVM (SUN)
10) GMM-SVM T-normed, without unsupervised adaptation (TNO)
11) GMM-SVM T-normed, with unsupervised adaptation (TNO)

For the non-adaptive variant STBU-1, indicated as STBU-1N in this paper, we simply omitted the last sub-system.

The second submission, STBU-2, is the same as STBU-1 in all respects, except that the eigenchannel GMM sub-systems were omitted. This makes this STBU-2 a pure fusion of SVM sub-systems. The third submission, STBU-3 is the same as STBU-1, except that the non-linear calibration described in Section IV-D was added.

Table II describes results on the primary condition (English only trials) for development data (SRE-2005) and for evaluation data (SRE-2006). Results are reported for all sub-systems, together with fused results which are with (U) and without (N) unsupervised adaptation. Fig. 2 presents the results graphically, where curves for GMM, GMM-SVM, and MLLR-SVM are grouped to keep the legend size manageable. Note how the curves for SRE-2006 are rotated clockwise w.r.t. the curves for SRE-2005. The little cusps in the MLLR-SVM curves are a side-effect of the zero-insertions discussed in Section IV-B.

Table III describes results on all trials from development and evaluation data. Only results of the best sub-system from each category is presented. Fig. 3 presents the results graphically with the same grouping of individual systems.

A comparison of the calibration performances of STBU-1 versus STBU-3 is given in table IV, as measured[20] on all 2006 1conv4w-1conv4w trials (without unsupervised adaptation). The *fixed-operating-point* calibration performance can be judged by the discrepancy between $C_{\rm det}$ and $C_{\rm det}^{\rm min}$, indeed as planned, STBU-1 performed better than STBU-3. Conversely, the *general* calibration as judged by the discrepancy between $C_{\rm llr}$ and $C_{\rm llr}^{\rm min}$ shows STBU-3, described in Section IV-D, to be better than STBU-1.

Although the calibration performance of the STBU system was good enough to make it competitive with the other submissions in the NIST SRE-2006, we note that the calibration performance in this evaluation was somewhat poorer for most participants as compared to the 2005 and 2004 evaluations. It is unlikely that this problem can be solved within the fusion and calibration paradigm presented here. Rather one may have to improve the sub-systems and make them more robust against changes in the nature of the speech data.

[20]Recall sub-systems were developed on 2004 and earlier data, fusion and calibration was trained on 2005 data, and this test was performed on new unseen 2006 data.

TABLE II
RESULTS OF THE SUB-SYSTEMS AND THE SUBMITTED ONE ON PRIMARY CONDITION: ENGLISH TRIALS.

system	SRE-2005 data		SRE-2006 data		
	$C_{\rm det}^{\rm min}$	EER	$C_{\rm det}^{\rm min}$	EER	$C_{\rm det}$
GMM (BUT)	.0174	3.88%	.0178	3.44%	
GMM T-norm (BUT)	.0170	4.27%	.0159	3.44%	
GMM-SVM (SUN)	.0153	4.19%	.0171	3.61%	
GMM-SVM (BUT)	.0158	4.66%	.0185	3.71%	
GMM-SVM-U (TNO)	.0116	3.72%	.0185	3.81%	
GMM-SVM (TNO)	.0178	5.17%	.0190	4.10%	
GMM-SVM For (SDV)	.0221	6.05%	.0227	4.91%	
GMM-SVM Rev (SDV)	.0220	6.10%	.0238	5.18%	
MLLR3-SVM (SUN)	.0212	6.05%	.0218	4.49%	
MLLR3-SVM (BUT)	.0196	6.17%	.0220	4.78%	
MLLR2-SVM (SUN)	.0264	7.50%	.0270	5.56%	
STBU-1U	.0070	2.98%	.0132	2.26%	**0.0154**
STBU-1N	.0096	3.21%	.0126	2.32%	0.0155
STBU-2U	.0073	3.17%	.0147	3.07%	0.0210
STBU-2N	.0099	3.59%	.0147	3.07%	0.0210
STBU-3U			.0132	2.27%	0.0161
STBU-3N			.0126	2.32%	0.0160

TABLE III
THE BEST PERFORMING SUB-SYSTEMS FROM EACH CATEGORY AND THE SUBMITTED RESULTS ON ALL TRIALS.

system	SRE-2005 data		SRE-2006 data		
	$C_{\rm det}^{\rm min}$	EER	$C_{\rm det}^{\rm min}$	EER	$C_{\rm det}$
GMM (BUT)	.0201	4.83%	.0283	5.40%	
GMM-SVM (TNO)	.0192	5.77%	.0285	6.04%	
MLLR-SVM (BUT)	.0224	7.15%	.0327	7.57%	
STBU-1U	.0085	3.50%	.0208	3.30%	0.0249
STBU-1	.0114	3.97%	.0214	3.83%	0.0263

C. Unsupervised adaptation

Unsupervised adaptation is an 'operating mode' of processing the NIST speaker recognition trials. In this mode, the available speech for a particular trial is extended with all earlier speech trials that include the same speaker model as the current trial. The trial index files are built such that (target) test segments are ordered by recording date for the same model speaker. The operating mode was proposed by Claude Barras [34] at the SRE-2003 workshop, adopted in the following NIST SRE plan as an optional mode, analysed separately. The rationale for this mode was that for certain applications, such as access authentication, there will typically be many target trials available which can provide the system with more speech of the target speaker so that better models can be formed [35].

For reasons which we will discuss below, successful application in a NIST SRE is hard [36], [37], but it finally succeeded in SRE-2005 [36]. Although, that year, only one participant had attempted to run the unsupervised adaptation mode, it still was considered an interesting research area, so it was decided that in SRE-2006, unsupervised adaptation mode

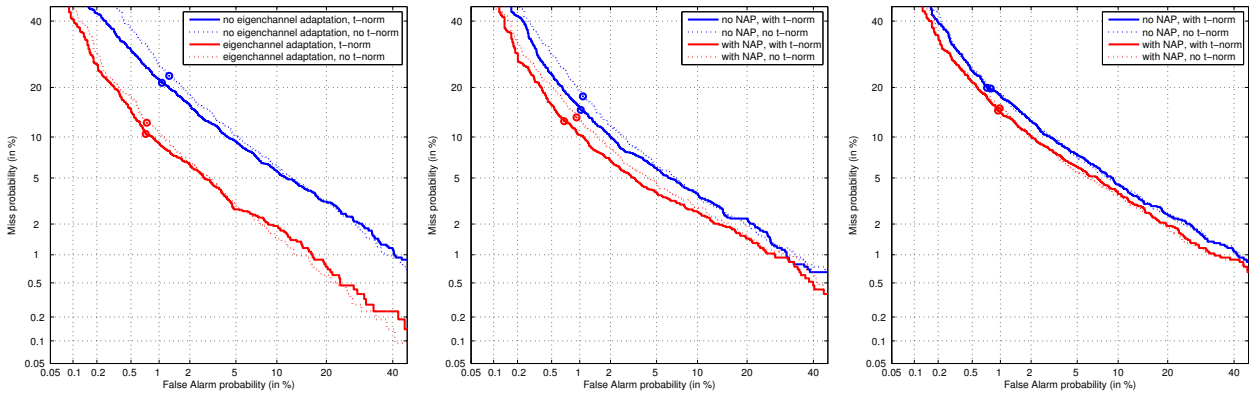


Fig. 1. Comparison of improvement with eigenchannel adaptation in the GMM system (left), NAP in GMM-SVM (middle) and NAP in MLLR-SVM (right). Results from SRE-2006, English-only trials. Circles indicate the $C_{\text{det}}^{\text{min}}$ operating point.

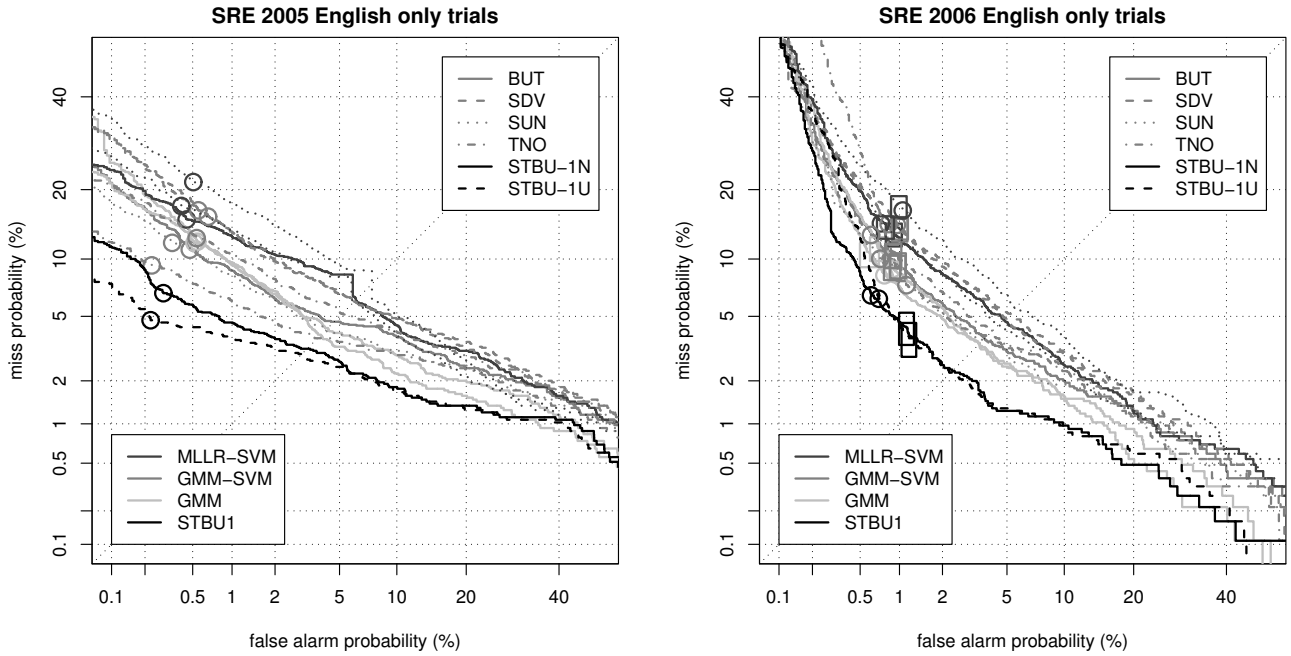


Fig. 2. DET curves for individual and merged systems, English only trials. MLLR-SVM, GMM-SVM and GMM sub-systems are grouped by colour. Left panel shows results on SRE-2005 data, where circles indicate $C_{\text{det}}^{\text{min}}$ operating point. Right panel shows SRE-2006 results, with additional boxes indicating 95 % confidence interval around the C_{det} operating point, based on calibration with SRE-2005 trials. The line type shows the site origin. STBU fusion results are in black, with a dashed curve for the unsupervised adaptation mode.

TABLE IV
COMPARISON OF CALIBRATION OF STBU-1 VS STBU-3, SRE-2006 ALL TRIALS.

SYSTEM	C_{llr}	$C_{\text{llr}}^{\text{min}}$	C_{det}	$C_{\text{det}}^{\text{min}}$
STBU-1	0.198	0.152	0.0263	0.0214
STBU-3	0.188	0.152	0.0274	0.0214

results could be entered as *primary system*.

There are different approaches to performing unsupervised adaptation, ranging from simple threshold-based inclusion of the test segment as extra training to score-weighted adaptation of the current model [35], [34], [37], [38], but all of them depend on proper calibration of the scores. This means that the *calibration will influence the position and shape* of the DET curve, as well as C_{det} and C_{llr}. Further, as has been pointed out earlier [34], [36], the *evaluation priors* of target and non-target trials, as well as the number of target-trials for each model speaker determine the potential success of application of unsupervised adaptation. This is different from the 'normal mode' of operation, where the evaluation priors do not determine the performance measures such as C_{det} and EER. A last major difference between the two operating modes is the influence of 'pathological data' in the evaluation. In the much appreciated data collection efforts and quality control it is inevitable, given the large amount of trials in evaluations (over 50 000 in SRE-2006), that there are speech files which contain little or no speech, are duplicates, or have the wrong language or speaker ID associated with it. For the 'normal mode' of operation this causes little problems, because in a standard post evaluation quality control procedure by NIST,

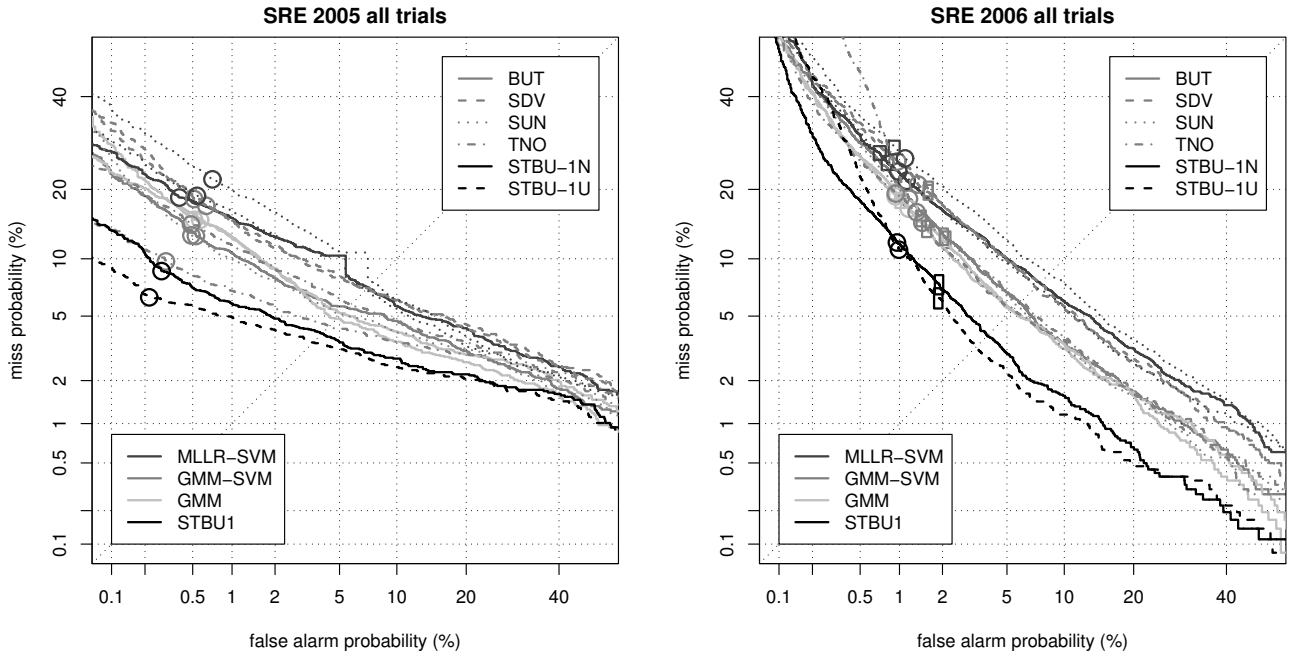


Fig. 3. DET curves for individual and merged systems, all trials. Colours, symbols and line type are the same as for Fig. 2.

trials involving these pathological files are discarded from further analysis. However, for the unsupervised adaptation mode, these pathological speech files can cause a major problem because the adaptive speaker model may deteriorate if such a file is not properly detected.

One sub-system (TNO) applied a simple adaptation scheme. It is based on earlier work [36] and extended to include the GMM-SVM-NAP technology. Basically, for each trial, the T-normed score s is calculated. If s exceeds a predetermined threshold a, the speech data in the test segment is used to MAP adapt the means in the GMM for the current model speaker, using a relevance factor r. The new means are used to build a new SVM, which is used for subsequent trials. The results for the development test (SRE-2005) and evaluation are summarized in Table V, and the DET-curves are shown in Fig. 4. Note, that these are the results of only one sub-system of the STBU submission. Qualitatively, the adaptation results are similar for the total system, but the effects are less pronounced due to the importance of several other sub-systems.

We tuned the parameters $a = 4$ and $r = 36$ to obtain optimum $C_{\rm det}^{\rm min}$ for SRE-2005, and applied these to SRE-2006. A speech file was classified as 'potentially pathological' if either the range of frame energy did not exceed 30 dB (assuming the file contains no speech) or if the SVM score, before T-norming, exceeded 0.95 (an assumed copy of a speech segment). For these trials, no adaptation was carried out. As it turns out, none of these trials survived the post evaluation quality control of NIST.

As can be observed from the table and the DET-curves, the discrimination performance increased dramatically for the development test (34 % relative drop in $C_{\rm det}$), but hardly at all for the evaluation (6 % relative drop in $C_{\rm det}$). The 'knee' close to the decision operating point for SRE-2006 is typical of runs where adaptation has been applied too aggressively (low a and r). It shows the effect of 'false adaptations' which spoil a speaker model and lead to over-optimistic scores for subsequent non-target trials. The 53966 trials in SRE-2006 lead to 5003 adaptations, of which 61.4 % were correct, 13.9 % false adaptations, and 24.7 % unknown, because these trials were later removed from the official scoring by NIST due to the various problems described earlier. Even though 'only' 2518 trials were removed from the original trial index file, 1223 of these (49 %) were used for adaptation of speaker models. On the other hand, of a potential 3612 target trials, only 14.9 % were missed for adaptation (see Table V).

As a post-evaluation experiment, 'post1,' we ran our adaptive mode parameters on the list of trials that were kept after the post evaluation quality control. Oddly enough, we observe from Table V that the performance *decreases* under this condition. Apparently, the 'pathological files' that plagued so many researchers during the evaluation, helped our sub-system in unsupervised adaptation mode. Perhaps some speakers who had enrolled twice under a different identity in the data collection process, and whose 'non-target trials' were later removed, actually helped in adaptation mode.

We attribute the poor adaptation performance to the high probability of False Adaptation [34], which is an order of magnitude larger than in the development test. This is not only due to miscalibration, but also because the DET-curve has a steeper slope. Indeed, optimizing the threshold as a post-evaluation experiment 'post2' to $a = 5$ leads to the expected larger benefit of unsupervised adaptation (21.5 % drop in $C_{\rm det}$), with a much lower False Adaptation probability.

D. Language dependence

We have observed that the performance in the primary condition (English only trials, Table II), is much better than

TABLE V
PERFORMANCE MEASURES FOR THE TNO SUB-SYSTEM IN NORMAL AN UNSUPERVISED ADAPTATION MODES, FOR DEVELOPMENT TEST (NO CALIBRATION, FIXED THRESHOLD OF 3), EVALUATION AND POST-EVALUATION EXPERIMENT. ALL (POST QUALITY CONTROL) TRIALS ARE INCLUDED. TWO POST-EVALUATION EXPERIMENTS ARE INCLUDED AS WELL. THE LAST TWO COLUMNS INDICATE THE PROBABILITY OF FALSE ADAPTATION AND MISSED ADAPTATION, RESPECTIVELY.

Mode	dataset	C_{det}	$C_{\mathrm{det}}^{\mathrm{min}}$	EER	C_{llr}	$C_{\mathrm{llr}}^{\mathrm{min}}$	$P_{\mathrm{FalseAd.}}$	$P_{\mathrm{missAd.}}$
Normal	SRE-2006	0.0335	0.0286	6.04 %	0.262	0.220		
Adapt.	SRE-2006	0.0315	0.0290	5.48 %	0.264	0.219	13.9 %	14.9 %
Normal	SRE-2005	0.0198	0.0189	5.79 %	0.629	0.220		
Adapt.	SRE-2005	0.0130	0.0124	4.38 %	0.572	0.171	1.1 %	13.8 %
Adapt. post1	SRE-2006	0.0349	0.0316	6.06 %	0.284	0.236	17.2 %	20.0 %
Adapt. post2	SRE-2006	0.0262	0.0227	4.73 %	0.220	0.182	5.5 %	25.4 %

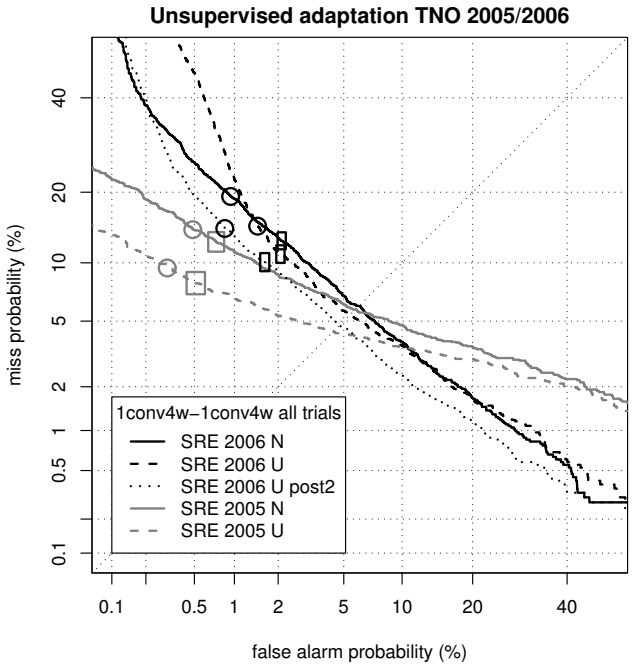


Fig. 4. DET-curves for the TNO sub-system in normal (solid lines) and unsupervised adaptation (dashed lines) modes, for evaluation and development test. Also included is a post-evaluation run with a more optimal threshold value (post2).

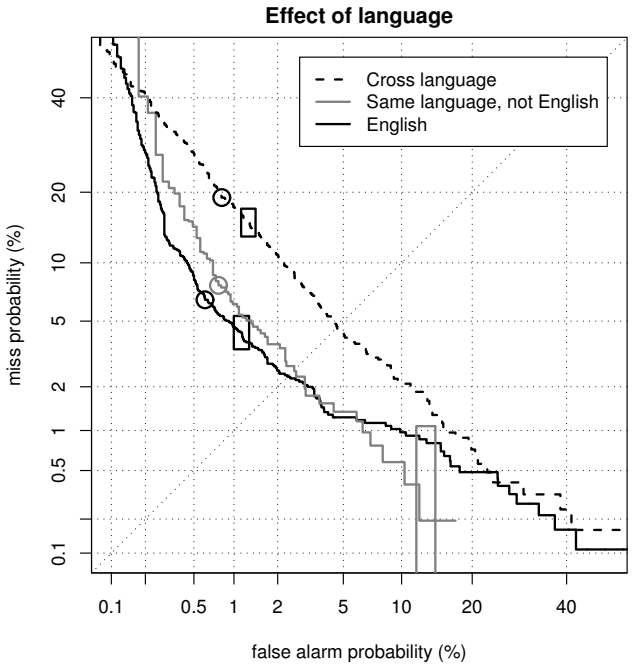


Fig. 5. DET-plots of the three different language condition analyzed in Table VI. The rectangle indicates the 95 % confidence interval around the decision point.

that of the entire evaluation (all trials, Table III). In this section we will analyse some language effects. A language dependence may be introduced by several parts of the system: the UBM, channel compensation, SVM background, score normalization and calibration. We split all valid trials of SRE-2006 into three conditions: *Same language English*, *Same language non-English* and *Cross language*. Note that by design of the evaluation, all cross-language trials involve English as one of the two spoken languages. In Table VI we summarize the important statistics of the three conditions.

Despite the low number of trials available for the non-English same-language condition, we can observe the following. The *discrimination* potential of the system seems similar for English and non-English same-language conditions, judged from a very similar EER, $C_{\mathrm{det}}^{\mathrm{min}}$ and $C_{\mathrm{llr}}^{\mathrm{min}}$. But the *calibration* for non-English trials is very poor (C_{det}, C_{llr}), compared to the English trials. This result suggests that the UBM and channel compensation components are less language dependent, but that there is a possible language dependence in score normalization and definitely in the calibration. Most sub-systems applied T-norm score normalization [39]. Because we applied predominantly English T-norm model speakers, we can imagine that non-English test segments will have lower scores for the T-norm models than the English test segments. This would lead to higher T-normed scores for non-English trials, for both target and non-target, such that the calibration is skewed towards more false alarms. Indeed, this is what is observed in Fig. 5.

A genuine discrimination loss is observed in the *cross language* trials. Interestingly, the calibration of the cross-language condition seems to be reasonable. This may be due to the fact that all cross-language target trials had English as one of the two speech segment languages. Apparently, having at least one English speech segment helps the calibration a lot.

All the described effects are qualitatively the same as observed for just a single sub-system (TNO) of the STBU fusion.

TABLE VI
LANGUAGE DEPENDENCE OF THE STBU-1 SYSTEM, FOR ENGLISH SAME-LANGUAGE TRIALS, NON-ENGLISH SAME LANGUAGE TRIALS AND CROSS LANGUAGE TRIALS.

Language	C_{det}	$C_{\mathrm{det}}^{\mathrm{min}}$	EER	C_{llr}	$C_{\mathrm{llr}}^{\mathrm{min}}$	N_{tar}	N_{non}
English	0.0155	0.0126	2.32 %	0.148	0.101	1854	22159
Non-English	0.128	0.0154	2.54 %	0.721	0.099	516	2857
Cross language	0.0277	0.0272	4,60 %	0.199	0.180	1242	22820

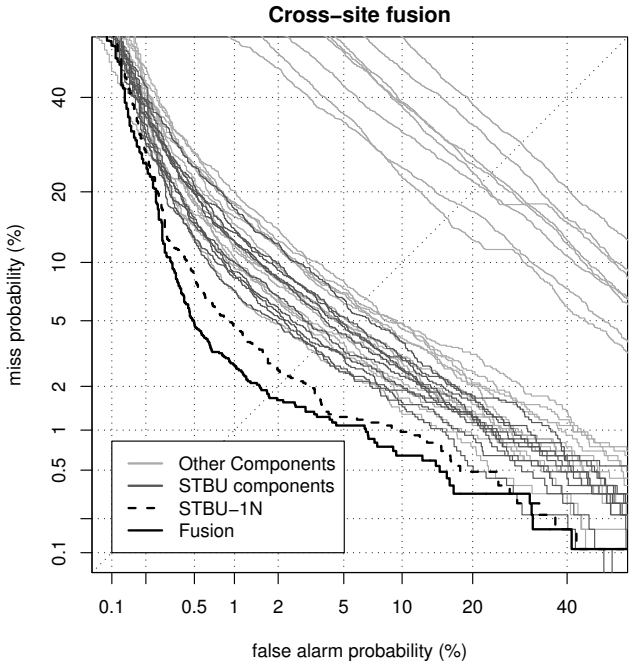


Fig. 6. Cross-site fusion: DET curves for individual and fused systems, for the English only trials condition of NIST SRE-2006.

VI. CROSS-SITE FUSION

As a final demonstration of the power of fusing diverse sub-systems, we increased the diversity and tripled the number of sub-systems by also including sub-systems, that performed well, from 6 other participating SRE-2006 sites. Together with 10 of the STBU sub-systems, this gave a total of 31 sub-systems, all non-adaptive. The fusion was trained on the supervised scores of all 1conv4w-1conv4w trials of SRE-2005 and then tested on the English 1conv4w-1conv4w trials of SRE-2006. See the DET-curves of Fig. 6, which shows (i) all 31 sub-systems, (ii) the original STBU-1 fusion, and (iii) the total fusion of all 31 systems. It is clear that the two fusions outperform any individual system, and that the bigger fusion (EER $= 1.7\,\%$) outperforms the original STBU fusion (EER $= 2.3\,\%$).

VII. CONCLUSION

The STBU system has demonstrated a few important principles that were exploited in reaching state-of-the-art speaker detection performance. (i) GMMs and SVMs are still important basic workhorses in speaker recognition, but alternative strategies like MLLR-SVM are not to be ignored. (ii) An abundance of suitable development data is perhaps the most important resource. Without the SRE-2004 and SRE-2005 databases, developing, testing and calibrating the powerful subspace channel compensation would not have been possible. Until recently, speaker recognition had been all about training individual speaker models. The emphasis has now shifted to the data-driven training of methods that can discriminate between speakers—we are no longer just training speaker models in isolation, each on a few minutes of speech. We are now training whole systems on the hundreds of hours of speech in whole NIST SRE databases. This is exemplified not only by eigenchannel and NAP, but also by fusion, which likewise needs to be trained on entire SRE databases. (iii) Calibration, in order to make *actual* decisions, has always been important in the NIST evaluations, but this had previously been measured only at the same fixed C_{det} operating point. The introduction of C_{llr} has now widened the scope of the calibration challenge, and so far not only the STBU system, but several other SRE-2006 participants have met this challenge successfully.

Despite these successes, several problem areas remain. As our investigation into the influence of the spoken language in detection performance shows, there is a strong effect on our system's calibration if trials are not English, and there is a reduction in discrimination if the segments of the trials are spoken in different languages. Perhaps these issues can be resolved with techniques similar to the channel compensation approaches. The unsupervised adaptation mode of processing trials did not deliver the large benefit we had expected, and we attribute this to a calibration mismatch, to which adaptation is very sensitive, and to the mysterious clockwise rotation of the DET curve observed for all systems that perform well. There remains the unsolved question of why new data collections and acoustic conditions seem to have an effect of rotation of the DET-curve—maybe to a more 'natural' state of equal width target and non-target score distributions. Continuing data collections, evaluations and research may on the long term provide us with an answer.

VIII. ACKNOWLEDGEMENTS

We would like to thank all SRE-2006 participants who provided their system results for Section VI, and granted permission to publish them anonymously. This work was partly supported by by European projects AMIDA (IST-033812) and Caretaker (FP6-027231), by Grant Agency of Czech Republic under projects No. 102/05/0278 and GP102/06/383, and by Czech Ministry of Education under project No. MSM0021630528. It is further partly supported by the project MultimediaN (http://www.multimedian.nl). MultimediaN is sponsored by the Dutch government under contract BSIK 03031.

REFERENCES

[1] P. Matějka, L. Burget, P. Schwarz, O. Glembek, M. Karafiát, J. Černocký, D. A. van Leeuwen, N. Brümmer, A. Strasheim, and F. Grézl, "STBU system for the NIST 2006 speaker recognition evaluation," in *Proc. ICASSP*, 2007, accepted for publication.

[2] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225–254, 2000.

[3] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten, "NIST and TNO-NFI evaluations of automatic speaker recognition," *Computer Speech and Language*, vol. 20, pp. 128–158, 2006.

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

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

[6] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. ICASSP*, 2004, pp. 37–40.

[7] A. Solomonoff, W. Campbell, and I. BoardmanCampbell, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP*, vol. I, Philadelphia, PA, USA, Mar. 2005, pp. 629–632.

[8] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluation chronicles—part 2," in *Proc. Odyssey 2006 Speaker and Language recognition workshop*, San Juan, June 2006.

[9] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing, special issue on Robust Speech Recognition*, vol. 2, no. 4, pp. 578–589, 1994.

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

[11] N. Kumar, "Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition," Ph.D. dissertation, John Hopkins University, Baltimore, 1997.

[12] L. Burget, "Complementarity of speech recognition systems and system combination," Ph.D. dissertation, Brno University of Technology, Czech Republic, 2004.

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

[14] L. Burget, P. Matějka, O. Glembek, P. Schwarz, and J. H. Černocký, "Analysis of feature extraction and channel compensation in GMM speaker recognition system," *IEEE Trans. on Audio, Speech and Language Processing*, 2007, accepted.

[15] M. Mason, R. Vogt, B. Baker, and S. Sridharan, "Data-driven clustering for blind feature mapping in speaker verification," in *Proc. Eurospeech*, Lisbon, Portugal, Sep. 2005, pp. 3109–3112.

[16] N. Brümmer, "Spescom DataVoice NIST 2004 system description," in *Proc. NIST Speaker Recognition Evaluation 2004*, Toledo, Spain, Jun. 2004.

[17] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, May 2007.

[18] ——, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, May 2007.

[19] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independent speaker verification," in *Proc. Interspeech*, 2005, pp. 3117–3120.

[20] R. Vogt and S. Sridharan, "Experiments in session variability modelling for speaker verification," in *Proc. ICASSP*, vol. 1, Toulouse, France, May 2006, pp. 897–900.

[21] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori esitimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 291–298, 1994.

[22] R. Auckenthaler, M. Carey, and H. Lloyd-Tomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

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

[24] W. Campbell, "A SVM/HMM system for speaker recognition," in *Proc. ICASSP*, Hong Kong, Apr. 2003, pp. 156–159.

[25] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Proc. Eurospeech*, Lisbon, Portugal, Sep. 2005, pp. 2425–2428.

[26] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The 2005 AMI system for the transcription of speech in meetings," in *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2006, vol. 3869, pp. 450–462.

[27] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *Proc. ICASSP*. Toulouse: IEEE, 2006, pp. 97–100.

[28] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," http://www.csie.ntu.edu.tw/∼cjlin/libsvm, 2001.

[29] S. Pigeon, P. Druyts, , and P. Verlinde, "Applying logistic regression to the fusion of the nist'99 1-speaker submissions," *Digital Signal Processing*, 2000.

[30] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Stellenbosch University, 2007.

[31] D. W. Hosner and S. Lemeshow, *Applied Logistic Regression*. John Wiley & Sons, 1989.

[32] D. A. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification*, ser. Lecture Notes in Computer Science / Artificial Intelligence, C. Müller, Ed. Heidelberg - New York - Berlin: Springer, 2007, vol. 4343.

[33] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, pp. 230–275, 2006.

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

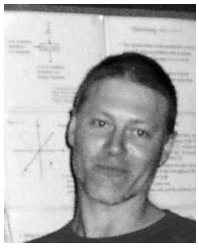
[35] N. Mirghafori and L. Heck, "An adaptive speaker verification system with speaker dependent a proiri deciasion thresholds," in *Proc. ICSLP*, 2002, pp. 589–592.

[36] D. A. van Leeuwen, "Speaker adaptation in the NIST speaker recognition evaluation 2004." in *Proc. Eurospeech*, 2005, pp. 1981–1984.

[37] E. G. Hansen, R. E. Slyh, and T. R. Anderson, "Supervised and unsupervised speaker adaptation in the nist 2005 speaker recognition evaluation," in *Proc. Odyssey 2006 Speaker and Language Recognition Workshop*, 2006.

[38] S.-C. Yin, P. Kenny, and R. Rose, "Speaker adaptation for factor analysis based speaker verification," in *Proc. Odyssey 2006 Speaker and Language recognition workshop*, San Juan, June 2006.

[39] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.



Niko Brümmer (M.Eng, University of Stellenbosch, 1988) is planning to submit his thesis entitled "Measuring, refining and calibrating speaker and language information extracted from speech," for a Ph.D, University of Stellenbosch in 2007. He has been employed as research engineer by Spescom DataVoice in South Africa, from 1990 to the present, on behalf of whom he has participated in 5 NIST Speaker Recognition Evaluations between 2000 and 2006, and also the NIST Language Recognition Evaluation 2005. His research interests include speaker and language recognition and the evaluation and improvement of pattern-recognition and machine-learning technologies via information theory.

Lukáš Burget (Ing. [MS]. Brno University of Technology, 1999, Ph.D. Brno University of Technology, 2004) is employed as assistant professor at at Faculty of Information Technology, University of Technology, Brno, Czech Republic. The topic of his PhD dissertation, that he successfully defended in November 2004, was: "Complementarity of Speech Recognition Systems an System combination". From 2000 to 2002, he was a visiting researcher at OGI Portland, USA under supervision of Prof. Hynek Hermansky. He is member of IEEE and ISCA. His scientific interests are in the field of speech processing, namely acoustic modeling for speech recognition.

Jan "Honza" Černocký (Ing. [MS] 1993 Brno University of Technology (BUT); Dr. [PhD] 1998 Université Paris XI and BUT) was with the Institute of Radio-electronics, BUT (Faculty of Electrical Engineering and Computer Science) as assistant professor from 1997. Since February 2002, he is with the Faculty of Information Technology (FIT), BUT as Associate Professor (Doc.) and Deputy Head of the Institute of Computer Graphics and Multimedia. With Prof. Hynek Hermansky he is leading the Speech@FIT group at FIT VUT. He supervises several PhD students, and coordinates Speech@FIT activities in several European and national projects. His research interests include signal processing, speech processing (very low bit rate coding, verification, recognition), segmental methods, data-driven determination of speech units and speech corpora. He is a member of IEEE and ISCA and serves on the board of Czechoslovak section of IEEE.

Ondřej Glembek (Ing. [MS]. Brno University of Technology, 2005) was student at Brno University of Technology, faculty of Electrical Engineering and Computer, later Faculty of Information Technology from 1999. From September till December 2003, he was at University of Joensuu, Finland as a participant of the Socrates/Erasmus program. From October till November 2004, he was working on a project concerning wavelet transforms at Izhevsk State Technical University, Izhevsk, Russia. From 2005, he is PhD student in Speech@FIT - he is concentrating on acoustic modeling for speech recognition, recognition of Czech and STK toolkit development.

František Grézl (Ing. [MS]. Brno University of Technology, 2000) is post-gradual student of Speech processing group at the Faculty of Information Technology (FIT), BUT since September 2000 and is planning to submit his doctoral thesis "Acoustic modeling for speech recognition" in summer 2007. He has been with the Anthropic speech processing group of Oregon Graduate Institute of Science and Technology, USA, with speech processing group at IDIAP research institute, Martigny, Switzerland and with ICSI International Computer Science Institute Berkeley, California under the AMI training programme. His main research interests include robust speech recognition and feature extraction.

Martin Karafiát (Ing. [MS]. Brno University of Technology, 2001) is post-gradual student in Speech@FIT at the Faculty of Information Technology (FIT), BUT since September 2001 and is planning to submit his doctoral thesis in autumn 2007. He was twice in the internship at University of Sheffield with Speech and Hearing Group, UK. Main reason for both internships was work on Large Vocabulary Continuous Speech Recognizers for two EU-projects M4 (Multimodal Meeting manager) and AMI (Augmented multiparty interaction). His research interest is speech recognition—especially speech recognition with large vocabulary, including feature transforms and novel feature extractions such as TRAPs.

David A. van Leeuwen (Ir. [MS] 1984 Delft University of Technology, Dr. [PhD] 1993 University of Leiden) is with TNO Human Factors since 1994. He has been active in the field of large vocabulary continuous speech recognition (evaluation, development of Dutch system), word spotting, and speaker and language recognition. He has organized several benchmark evaluations (LVCSR Fr/Ge/BrEng: EU SQALE in 1995, Forensic Speaker Recognition NFI-TNO in 2003, LVCSR Dutch: N-Best in 2008). He has participated in NIST SRE, LRE and RT evaluations since 2003. He has been a representative in several NATO IST Research task groups on speech technology since 2002, and an ISCA member since 1995.

Pavel Matějka (Ing. [MS]. Brno University of Technology, 2001) is PhD student at Institute of Radioelectronics, Faculty of Electrical Engineering and Communication and Department of Computer Graphics and Multimedia, FIT, BUT. He is planning to submit his doctoral thesis "Language identification based on phonetic cues" in summer 2007. He has been with the Anthropic speech processing group at Oregon Graduate Institute of Science and Technology, USA. He is member of IEEE and ISCA. His research interests include speaker recognition, language identification, speech recognition, namely phoneme recognition based on novel feature extractions (temporal patterns) and neural networks. He has been active also in keyword-spotting and on-line implementation of speech processing algorithms. He was finalist in Student paper contest at ICASSP2006 in Toulouse.

Petr Schwarz (Ing. [MS]. Brno University of Technology, 2001) is post-gradual student of Speech processing group at the Faculty of Information Technology (FIT), BUT since September 2001 and is planning to submit his doctoral thesis "Robust phoneme recognition" in 2007. He has been with the Anthropic speech processing group of Oregon Graduate Institute of Science and Technology, USA. He is member of IEEE and ISCA. His research interests include speech recognition, namely phoneme recognition based on novel feature extractions (temporal patterns) and neural networks. He has been active also in keyword-spotting and on-line implementation of speech processing algorithms.

Albert Strasheim (B.Sc, University of Stellenbosch, 2003, B.Eng, University of Stellenbosch, 2005) is a Masters student at the University of Stellenbosch, Digital Signal Processing Lab, supervised by Prof. Johan du Preez. His research interests include software engineering and parallel and distributed systems, specifically their application to pattern recognition and machine learning problems.